

Amy Neustein (Ed.)

Text Mining of Web-Based Medical Content

Speech Technology and Text Mining in Medicine and Health Care



Series Editor
Amy Neustein

Additional Titles in the Series

Neustein (Ed.), *Speech and Automata in Health Care*
(forthcoming, November 2014), ISBN: 978-1-61451-709-2

Patil and Kulshreshtha (Eds.), *Signal and Acoustic Modeling for Speech and Communication Disorders* (forthcoming, May 2015), ISBN: 978-1-61451-759-7

Ganchev, *Computational Bioacoustics*
(forthcoming, May 2015), ISBN: 978-1-61451-729-0

Beals, Dahl, and Linebarger, *Speech and Language Technology for Language Disorders* (forthcoming, August 2015), ISBN: 978-1-61451-758-0

Text Mining of Web-Based Medical Content



Edited by
Amy Neustein

DE GRUYTER

Editor

Amy Neustein
800 Palisade Avenue
Suite 1809
Fort Lee, NJ 07024
USA
amy.neustein@verizon.net

ISBN 978-1-61451-541-8

e-ISBN (PDF) 978-1-61451-390-2

e-ISBN (epub) 978-1-61451-976-8

Library of Congress Cataloging-in-Publication data

A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.dnb.de>.

© 2014 Walter de Gruyter Inc., Boston/Berlin

Typesetting: Compuscript Ltd., Shannon, Ireland

Printing and binding: CPI Books GmbH, Leck

Cover image: MEHAU KULYK/SCIENCE PHOTO LIBRARY / Agentur Focus

∞ Printed on acid-free paper

Printed in Germany

www.degruyter.com

Preface

Text Mining of Web-Based Medical Content brings together a talented group of researchers devoted to the study of how to derive high quality information from online data sources, ranging from biomedical literature, electronic health records, query search terms, social media posts and tweets, to general health information found on the web. Using some of the latest empirical methods of knowledge extraction, the authors show how online content, generated by both professionals and laypersons, can be mined for valuable information about disease processes, adverse drug reactions not captured during clinical trials, and tropical fever outbreaks. In this anthology the authors show how to perform information extraction on a hospital intranet and how to build a social media search engine to glean information about patients' own experiences in interacting with health-care professionals. In addition, some of the authors have studied ways to improve access to online health information for those who suffer from visual impairments, while others have studied the use of information extraction techniques in sifting through YouTube video descriptions and radiographic image data.

This book is divided into four sections:

The first section closely examines methods and techniques for mining biomedical literature and electronic health records.

The section opens with a comprehensive overview of the application of text mining to biomedical knowledge extraction, analyzing both clinical narratives and medical literature. The authors demonstrate how the four main phases of biomedical knowledge extraction using text mining (text gathering, text preprocessing, text analysis, and presentation) may be used to obtain relevant information from vast online databases of health science literature and patients' electronic health records. They present various text mining tools that have been developed in both university and commercial settings, as well as an in-depth analysis of the differences between clinical text found in electronic health records and biomedical text found in online journals, books, and conference papers.

In the following chapter, the authors focus exclusively on patients' electronic health records, showing how clinical natural language processing (NLP) can effectively unlock detailed patient information from clinical narratives stored in such records. This chapter introduces the state-of-the-art work in clinical NLP. Using medication information extraction as a use case, the authors describe different methods to build clinical NLP systems, including rule-based, machine learning-based, and hybrid approaches. Applications of medication information extraction systems, such as *pharmacovigilance* (post-market surveillance of drugs) research, are also discussed in this chapter.

The section is rounded out by a fascinating report on two prototypes for performing information extraction on both a hospital intranet and on the World Wide Web. The authors show how they apply ontology-based information extraction to unstructured natural language sources to help enable semantic search of health information. They propose a general architecture capable of handling both private and public data. Two of their novel systems that are based on this architecture are presented in their chapter. The first system, MedInX, is a Medical Information eXtraction system, which processes textual clinical discharge records, performing automatic and accurate mapping of free text reports onto a structured representation. MedInX is designed to be used by health professionals, and by hospital administrators and managers, allowing a search of the contents of its automatically populated ontologies. The second system, SPHInX, attempts to perform semantic search on health information publicly available on the web in Portuguese. The authors provide usage examples and evaluation results that show the potential of their proposed approach to performing information extraction on unstructured text found in hospital records and on the Internet.

The second section explores machine learning techniques for mining medical search queries and health-related social media posts and tweets. In so doing, the authors demonstrate a keen grasp of how laypersons use the web for seeking health information and reassurance. They focus both on search query data entered in the Google search engine and on the health-related user-generated content found on social media sites and on Twitter.

The section begins with a chapter titled “Predicting Dengue Incidence in Thailand from Online Search Queries that Include Weather and Climatic Variables.” The chapter presents machine learning techniques to help public health agencies mitigate vector borne disease, in particular dengue outbreaks. Search queries from digital sources are used to forecast the number of dengue cases prior to officially reported cases. This is achieved by processing query terms related to vector-borne dengue disease. Since climate has been correlated to the vector’s dynamics, query terms related to weather are utilized for the forecasting of dengue cases.

All in all, one can certainly see the value of mining search query data to predict the number of dengue cases so that public health authorities may devise adequate interventions to address dengue outbreaks before they reach catastrophic proportion.

The chapter on monitoring users’ query search terms in predicting a disease outbreak is followed by a fascinating chapter that addresses the other side of the coin. That is, in this subsequent chapter the authors provide a detailed study of how users sometimes divulge *too much* personal health information on line.

The authors opine that with the increasing amount of personal information that is shared on social networks, it is possible that the users might inadvertently reveal some personal health information that may have untoward consequences for the user. They show that personal health information can be detected and, if necessary, protected. They present empirical support for this hypothesis, by showing how two existing well-known electronic medical resources MedDRA and SNOMED help to detect personal health information (PHI) in messages retrieved from a social network site, MySpace. To do so, they introduce a new measure – Risk Factor of Personal Information – that assesses the likelihood that a term would reveal personal health information. They synthesize a profile of a potential PHI leak in a social network, and demonstrate that this task benefits from the emphasis on the MedDRA and SNOMED terms. Using machine learning techniques to validate the importance of terms detected by these two medical dictionaries, they show that their study findings are robust in detecting sentences and phrases that contain users' personal health information.

The section concludes with a thought-provoking analysis of the expanding role of social media for those who seek health information and for those who study social trends based on patient blog postings. Yet, the authors wisely point out that this new medium of communication has its limitations too. Namely, the current inability to access and curate relevant information in the ever-increasing gamut of messages. In their chapter, the authors demonstrate how they seek to understand and curate laypersons' personal experiences on Twitter. To do so, they propose some solutions to improve search, summarization, and visualization capabilities for Twitter (or social media in general), in both real time and retrospectively. In essence, they provide a basic recipe for building a search engine for social media and then make it increasingly more intelligent through smarter processing and personalization of search queries, tweet messages, and search results. In addition, they address the summarization aspect by visualizing topical clusters in tweets and further classifying the retrieval results into topical categories that serve professionals in their work. Finally, they discuss information curation by automating the classification of the information sources as well as combining, comparing, and correlating tweets with other sources of health information. In discussing all these important features of social media search engines they present systems, which they themselves have developed to help identify useful information in social media.

The third section presents speech and audio technologies for improving access to online content for the computer-illiterate and the visually impaired. The authors report on thoughtfully designed systems that help *democratize* the availability of online health information for those who cannot readily access this information on their own.

The section begins with an empirical study of user satisfaction with a health dialogue system designed for the Nigerian low-literate, computer-illiterate, and visually impaired. The author shows how this health dialogue system provides health information about lassa fever, malaria fever, typhoid fever and yellow fever to those who cannot access this information on line. The author points out that since this information on the Internet is mainly delivered in text format, it is only available to a small percentage of the population due to inadequate Internet access and the low level of literacy in Nigeria. The chapter reports on the development, acceptability, and user satisfaction with this dialogue system, which provides health information about these tropical fevers. The author conducted his cross-sectional study using a questionnaire that gathered demographic data about the study participants and their satisfaction and readiness to accept the health dialogue system. The user satisfaction results showed a mean of 3.98 (approximately 4), which is the recommended average for a good usability study. Dialogue systems of this kind help to provide cost-effective and equitable access to health information that can protect the population from tropical disease outbreaks. They serve the low-literate, the computer-illiterate, and the visually impaired.

The chapter on user satisfaction with a health dialogue system is followed by a fascinating presentation of the Smith-Kettlewell Eye Research Institute's Descriptive Video Exchange (DVX) project, which helps the blind and the visually impaired gain access to the information contained in health-related videos found on the web. The author shows step-by-step how DVX provides a framework that enables a large number of people, both amateur and professional, to create descriptions of video data both quickly and easily. The author shows how DVX distributes those descriptions so that they are available to anyone on the Internet and, in particular, provides a special service for the visually impaired. He points out that DVX when combined with speech recognition can greatly improve video search. In short, this chapter closely examines the use of audio (and text-to-speech) description, created through *crowd sourcing*, to improve video accessibility for the blind and the visually impaired.

The fourth section serves as the coda to this book. The contributors to this section have studied the use of information extraction techniques for accessing both medical images stored in digital libraries and health-related video material found on the web.

The section begins by taking a close look at information extraction from medical images. The authors present in detail their evaluation of a novel automatic image annotation system using semantic-based information retrieval. However, first they show the obstacles for image annotations, namely, the semantic gap problem – it is hard to extract semantically meaningful entities

when using low-level image features – and the lack of correspondence between the keywords and image regions in the training data. Then, they show that though content-based visual information retrieval (CBVIR) and image annotation has attracted a lot of interest, namely from the image engineering, computer vision, and database community, current methods of the CBVIR systems only focus on appearance-based similarity, i.e., the appearance of the retrieved images is similar to that of a query image. As a result, there is very little semantic information exploited.

To overcome this obstacle the authors have developed a semantic-based visual information retrieval (SBVIR) system while recognizing that two steps are required: (1) to extract the visual objects from images; and (2) to associate semantic information with each visual object. The authors show that the first step can be achieved by using segmentation methods applied to images, while the second step can be achieved by using semantic annotation methods applied to the visual objects extracted from images. They point out that for testing their annotation module they used a set of 2000 medical images: 1500 of images in the training set and 500 test images. For testing the quality of their segmentation algorithm they used a database consisting of 500 medical images of the digestive system that were captured by an endoscope. Their test results, based on looking at the assigned words to see if they were relevant to the image in question, have proven that their automatic image annotation system augurs well in the diagnostic and treatment process. The authors' novel approach to information extraction from medical images is no doubt a first step toward larger studies of automatic image annotation for indexing, retrieving, and understanding large collections of image data. Moreover, the field of information extraction, which is considered a subtask of text mining, can only benefit from such rigorous studies of multimedia document processing, involving automatic annotation and content extraction from medical images.

The book concludes with a study of video metadata by focusing on the title and description of health-related videos found on the web to see if a lay user in search of medical information can perform a successful online search.

The authors contend that though huge amounts of health-related videos are available on the Internet (and health consumers are increasingly looking for answers to their health problems and health concerns by searching for videos online), a critical factor in identifying relevant videos based on a textual query is the accuracy of the *metadata* with respect to video content. The authors focus on how reputable health videos providers, such as hospitals and health organizations, describe diabetes-related video content and the frequency with which they use standard terminology found in medical thesauri. Their study compared video title

and description to medical terms extracted from the MeSH and ICD-10 vocabularies, respectively. They found that only a small number of videos were described using medical terms (4% of the videos included an exact ICD-10 term; and 7% an exact MeSH term). Furthermore, of all those videos that used medical terms in their title/description, they found an astonishingly low variety of diabetes-related medical terms used. For example, the video titles and descriptions brought up only 2.4% of the ICD-10 terms and 4.3% of MeSH terms, respectively.

The authors make the point that these figures certainly give one pause to think as to how many useful health videos are haplessly eluding online patient search because of the sparse use of appropriate terms in video titles and descriptions. Though no one would deny that including medical terms in video title and description is useful to patients who are searching for relevant health information, the authors point out that by adopting good practices for titling and describing health-related videos it may serve another purpose as well. That is, they can help producers of YouTube videos to identify and address the gaps in the delivery of informational resources that patients need to be able to monitor their own health. The authors conclude that sadly, as the situation is now, neither patients nor producers of health videos are able to explore the collection of online materials in the same systematic manner as the medical professional explores medical domains using MEDLINE. The authors pose the question: Why can't we have the same level of rigorous and systematic curating of patient-related health videos as we have for other medical content on the web?

Perhaps this book will provide the answer.

Amy Neustein
Fort Lee, NJ
September, 2014

Contents

Preface — v

List of authors — xix

Part I Methods and techniques for mining biomedical literature and electronic health records — 1

Amy Neustein, S. Sagar Imambi, Mário Rodrigues, António Teixeira and Liliana Ferreira

1 Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature — 3

- 1.1 Introduction — 3
- 1.2 Background — 6
 - 1.2.1 Clinical and biomedical text — 6
 - 1.2.2 Information retrieval — 8
 - 1.2.2.1 Information retrieval process — 9
 - 1.2.3 Information extraction — 10
 - 1.2.4 Challenges to biomedical information extraction systems — 10
 - 1.2.5 Applications of biomedical information extraction tools — 12
- 1.3 Biomedical knowledge extraction using text mining — 13
 - 1.3.1 Unstructured text gathering and preprocessing — 14
 - 1.3.1.1 Text gathering — 14
 - 1.3.1.2 Text preprocessing — 15
 - 1.3.2 Extraction of features and semantic information — 15
 - 1.3.3 Analysis of annotated texts — 16
 - 1.3.3.1 Algorithms for text classification — 18
 - 1.3.3.2 Classification evaluation measures — 20
 - 1.3.4 Presentation — 23
- 1.4 Text mining tools — 23
- 1.5 Summary — 26
 - Appendix “A” — 27
 - References — 28

Hua Xu and Joshua C. Denny

2 Unlocking information in electronic health records using natural language processing: a case study in medication information extraction — 33

- 2.1 Introduction to clinical natural language processing — 33
- 2.2 Medication information in EHRs — 35

- 2.3 Medication information extraction systems and methods — 37
 - 2.3.1 Relevant work — 37
 - 2.3.2 Summary of approaches — 39
 - 2.3.2.1 Rule-based methods — 39
 - 2.3.2.2 Machine learning-based methods — 40
 - 2.3.2.3 Hybrid methods — 42
- 2.4 Uses of medication information extraction tools in clinical research — 42
- 2.5 Challenges and future work — 43
 - References — 44

António Teixeira, Liliana Ferreira and Mário Rodrigues

3 Online health information semantic search and exploration: reporting on two prototypes for performing information extraction on both a hospital intranet and the world wide web — 49

- 3.1 Introduction — 49
- 3.2 Background — 51
- 3.3 Related work — 52
 - 3.3.1 Semantic search — 53
 - 3.3.2 Health information search and exploration — 53
 - 3.3.3 Information extraction for health — 54
 - 3.3.4 Ontology-based information extraction – OBIE — 55
- 3.4 A general architecture for health search: handling both private and public content — 56
- 3.5 Two semantic search systems for health — 58
 - 3.5.1 MedInX — 58
 - 3.5.1.1 MedInX ontologies — 59
 - 3.5.1.2 MedInX system — 61
 - 3.5.1.3 Representative results — 62
 - 3.5.2 SPHInX – Semantic search of public health information in Portuguese — 64
 - 3.5.2.1 System architecture — 64
 - 3.5.2.2 Natural language processing — 65
 - 3.5.2.3 Semantic extraction models — 65
 - 3.5.2.4 Semantic extraction and integration — 66
 - 3.5.2.5 Search and exploration — 67
- 3.6 Conclusion — 69
 - Acknowledgments — 70
 - References — 70

Part II Machine learning techniques for mining medical search queries and health-related social media posts and tweets — 75

Jedsada Chartree, Angel Bravo-Salgado, Tamara Jimenez and Armin R. Mikler

4 Predicting dengue incidence in Thailand from online search queries that include weather and climatic variables — 77

- 4.1 Introduction — 77
- 4.1.1 Dengue disease in the world — 78
- 4.2 Epidemiology of dengue disease — 79
- 4.2.1 Temperature change and the ecology of *A. aegypti* — 80
- 4.3 Using online data to forecast incidence of dengue — 83
- 4.3.1 Background and related work — 83
- 4.3.2 Methodology for dengue cases prediction — 86
- 4.3.2.1 Framework — 86
- 4.3.2.2 Data sets — 87
- 4.3.2.3 Predictive models — 89
- 4.3.2.4 Validation — 92
- 4.3.3 Prediction analysis — 93
- 4.3.3.1 Multiple linear regression — 93
- 4.3.3.2 Artificial neural network — 96
- 4.3.3.3 Comparison of predictive models — 100
- 4.3.4 Discussion — 100
- 4.4 Conclusion — 102
- References — 103

Kambiz Ghazinour, Marina Sokolova and Stan Matwin

5 A study of personal health information posted online: using machine learning to validate the importance of the terms detected by MedDRA and SNOMED in revealing health information in social media — 107

- 5.1 Introduction — 107
- 5.2 Related background — 108
- 5.2.1 Personal health information in social networks — 108
- 5.2.2 Protection of personal health information — 111
- 5.2.3 Previous work — 112
- 5.3 Technology — 113
- 5.3.1 Data mining — 113
- 5.3.2 Machine learning — 114
- 5.3.3 Information extraction — 114
- 5.3.4 Natural language processing — 115
- 5.4 Electronic resources of medical terminology — 116

- 5.4.1 MedDRA and its use in text data mining — 116
- 5.4.2 SNOMED and its use in text data mining — 117
- 5.4.3 Benefits of using MedDRA and SNOMED — 119
- 5.5 Empirical study — 119
 - 5.5.1 MySpace data — 119
 - 5.5.2 Data annotation — 120
 - 5.5.3 MedDRA results — 121
 - 5.5.4 SNOMED results — 122
- 5.6 Risk factor of personal information — 123
 - 5.6.1 Introducing RFPI — 123
 - 5.6.2 Results from MedDRA and SNOMED — 124
 - 5.6.3 Challenges in detecting PHI — 126
- 5.7 Learning the profile of PHI disclosure — 127
 - 5.7.1 Part I – Standard bag of words model — 127
 - 5.7.2 Part II – Special treatment for medical terms — 128
- 5.8 Conclusion and future work — 128
 - Acknowledgment — 130
 - References — 130

Hanna Suominen, Leif Hanlen and Cécile Paris

- 6 Twitter for health – building a social media search engine to better understand and curate laypersons’ personal experiences — 133**
 - 6.1 Introduction — 133
 - 6.2 Background — 136
 - 6.2.1 Social media as a source of health information — 136
 - 6.2.2 Information search on social media — 138
 - 6.3 Proposed solutions — 141
 - 6.3.1 Tools for information retrieval on twitter — 141
 - 6.3.1.1 Basic recipe for building a search engine — 142
 - 6.3.1.2 Solutions — 143
 - 6.3.1.3 Health concerns, availability of clean water and food, and other information for crisis management knowledge from twitter — 148
 - 6.4 Background — 148
 - 6.5 Some solutions — 149
 - 6.6 Tools for combining, comparing, and correlating tweets with other sources of health information — 156
 - 6.7 Discussion — 160
 - 6.8 Related solutions — 160
 - 6.8.1 Maps applications for disease monitoring — 161
 - 6.8.2 Maps applications in crisis situations — 161

- 6.8.3 Extraction systems to monitor relationships between drugs and adverse events — 162
- 6.8.4 An early warning system to discover unrecognized adverse drug events — 164
- 6.9 Methods for information curation — 166
- 6.10 Future work — 167
 - Acknowledgments — 168
 - References — 168

Part III Using speech and audio technologies for improving access to online content for the computer-illiterate and the visually impaired — 175

Olufemi Oyelami

7 An empirical study of user satisfaction with a health dialogue system designed for the Nigerian low-literate, computer-illiterate, and visually impaired — 177

- 7.1 Introduction — 177
- 7.2 Related work — 178
- 7.3 Dialogue systems — 181
- 7.4 Methods — 182
 - 7.4.1 Participants — 182
 - 7.4.2 Demographics of the participants — 183
 - 7.4.3 Data collection — 183
 - 7.4.4 Data analysis — 183
- 7.5 Health dialogue system (HDS) — 183
- 7.6 Results — 184
 - 7.6.1 Experiences with mobile/computing devices — 184
 - 7.6.2 User satisfaction and acceptability of HDS — 186
- 7.7 Conclusion — 187
 - Acknowledgment — 187
 - References — 187

Keith M. Williams

8 DVX – the descriptive video exchange project: using crowd-based audio clips to improve online video access for the blind and the visually impaired — 191

- 8.1 Current problems with video data — 191
- 8.2 The description solution — 192
 - 8.2.1 What is description? — 192
 - 8.2.2 Description for the visually impaired — 192

- 8.2.2.1 Current types — 193
- 8.3 Architecture of DVX — 194
 - 8.3.1 The DVX server — 194
 - 8.3.1.1 Major data elements, attributes and actions — 195
 - 8.3.1.2 Current implementation — 198
 - 8.3.1.3 Tomcat servlet container — 198
 - 8.3.1.4 Applications — 199
- 8.4 DVX solves description problems — 204
- 8.5 DVX and video search — 205
- 8.6 Conclusion — 206
 - Acknowledgment — 206

Part IV Visual data: new methods and approaches to mining radiographic image data and video metadata — 207

Dumitru Dan Burdescu, Liana Stanescu and Marius Brezovan

9 Information extraction from medical images: evaluating a novel automatic image annotation system using semantic-based visual information retrieval — 209

- 9.1 Introduction — 210
- 9.2 Background — 211
- 9.3 Related work — 213
- 9.4 Architecture of system — 215
- 9.5 The segmentation algorithm – graph-based object detection (GBOD) — 219
- 9.6 Experimental results — 230
- 9.7 Conclusions — 235
 - References — 237

Randi Karlsen, Jose Enrique Borrás Morell, Johan Gustav Bellika and Vicente Traver Salcedo

10 Helping patients in performing online video search: evaluating the importance of medical terminology extracted from MeSH and ICD-10 in health video title and description — 241

- 10.1 Introduction — 242
- 10.2 Data and methods — 243
 - 10.2.1 Obtaining video data — 243
 - 10.2.2 Detecting medical terms in video title and/or description — 245
 - 10.2.3 Medical vocabularies — 246

10.3	Results — 247
10.3.1	ICD-10 Results — 247
10.3.2	MeSH Results — 250
10.3.3	Terms used in video titles and descriptions — 253
10.3.4	Occurrences of terms – when discarding the most common terms — 255
10.4	Discussion — 256
10.4.1	Findings — 256
10.4.2	How ICD-10 and MeSH terms can be useful — 257
10.4.3	Discriminating power of terms — 258
10.4.4	The uniqueness of our study when compared to other work — 258
10.5	Conclusion — 260
	Acknowledgments — 260
	References — 261

Editor's biography — 263

List of authors

Johan Gustav Bellika

Norwegian Centre for Integrated Care
and Telemedicine
Tromsø
Norway

Angel Bravo-Salgado

Center for Computational Epidemiology
and Response Analysis (CeCERA)
University of North Texas
Denton

Marius Brezovan

Faculty of Automation
Computers and Electronics
University of Craiova
Romania

Dumitru Dan Burdescu

Faculty of Automation
Computers and Electronics
University of Craiova
Romania

Jedsada Chartree

Center for Computational
Epidemiology and Response Analysis
(CeCERA)
University of North Texas
Sisaket Rajabhat University
Thailand

Joshua C. Denny

Department of Biomedical Informatics
and Medicine
Vanderbilt University
Nashville
Tennessee

Liliana Ferreira

Department of Electronics
Telecommunications and Informatics/
IEETA
University of Aveiro
Portugal

Kambiz Ghazinour

School of Electrical Engineering and
Computer Science
University of Ottawa
Canada

Leif Hanlen

NICTA
Faculty of Health
University of Canberra
College of Engineering and Computer
Science
Australian National University
Australia

S. Sagar Imambi

T.J.P.S. College
Guntur
India

Tamara Jimenez

Center for Computational Epidemiology
and Response Analysis (CeCERA)
University of North Texas
Denton

Randi Karlsen

Department of Computer Science
UiT The Arctic University of Norway
Tromsø
Norway

Stan Matwin

School of Electrical Engineering and
Computer Science
University of Ottawa
Canada
Institute for Big Data Analytics
Dalhousie University
Faculty of Computer Science
Dalhousie University
Halifax
Nova Scotia

Armin R. Mikler

Center for Computational
Epidemiology and Response
Analysis (CeCERA)
University of North Texas
Denton

Jose Enrique Borrás Morell

Department of Computer Science
UiT The Arctic University of Norway
Tromsø
Norway

Amy Neustein

Editor-in-Chief
International Journal of Speech
Technology
Series Editor of Speech Technology
and Text Mining in Medicine and
Health Care (De Gruyter)
Founder and CEO
Linguistic Technology Systems
Fort Lee
New Jersey

Olufemi Oyelami

Department of Computer and
Information Sciences
Covenant University
Ota
Nigeria

Cécile Paris

CSIRO
Computational Informatics
Macquarie University
Sydney
Australian National University
Canberra
Australia

Mário Rodrigues

ESTGA/IEETA
University of Aveiro
Portugal

Vicente Traver Salcedo

ITACA – Health and Wellbeing
Technologies
Universidad Politécnica de Valencia
Spain

Marina Sokolova

School of Electrical Engineering and
Computer Science
University of Ottawa
Canada
Faculty of Medicine
University of Ottawa
Institute for Big Data Analytics
Dalhousie University
Halifax
Nova Scotia

Liana Stanescu

Faculty of Automation
Computers and Electronics
University of Craiova
Romania

Hanna Suominen

NICTA
Faculty of Health
University of Canberra
College of Engineering and Computer
Science
Australian National University
Australia

António Teixeira

Department of Electronics
Telecommunications and Informatics/
IEETA
University of Aveiro
Portugal

Keith M. Williams

Senior Programmer Analyst
Smith-Kettlewell Eye Research Institute
San Francisco

Hua Xu

School of Biomedical Informatics
University of Texas Health Science Center
Houston

**Part I Methods and techniques for mining
biomedical literature and electronic
health records**

Amy Neustein, S. Sagar Imambi, Mário Rodrigues,
António Teixeira and Liliana Ferreira

1 Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature

Abstract: One of the tools that can aid researchers and clinicians in coping with the surfeit of biomedical information is text mining. In this chapter, we explore how text mining is used to perform biomedical knowledge extraction. By describing its main phases, we show how text mining can be used to obtain relevant information from vast online databases of health science literature and patients' electronic health records. In so doing, we describe the workings of the four phases of biomedical knowledge extraction using text mining (text gathering, text preprocessing, text analysis, and presentation) entailed in retrieval of the sought information with a high accuracy rate. The chapter also includes an in depth analysis of the differences between clinical text found in electronic health records and biomedical text found in online journals, books, and conference papers, as well as a presentation of various text mining tools that have been developed in both university and commercial settings.

1.1 Introduction

The corpus of biomedical information is growing very rapidly. New and useful results appear every day in research publications, from journal articles to book chapters to workshop and conference proceedings. Many of these publications are available online through journal citation databases such as Medline – a subset of the PubMed interface that enables access to Medline publications – which is among the largest and most well-known online databases for indexing professional literature. Such databases and their associated search engines contain important research work in the biological and medical domain, including recent findings pertaining to diseases, symptoms, and medications. Researchers widely agree that the ability to retrieve desired information is vital for making efficient use of the knowledge found in online databases. Yet, given the current state of *information overload* efficient retrieval of useful information may be severely hampered. Hence, a retrieval system “should not only be able to retrieve the

sought information, but also filter out irrelevant documents, while giving the relevant ones the highest ranking” (Ramampiaro 2010).

One of the tools that can aid researchers and clinicians in coping with the surfeit of information is text mining. Text mining refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Those in the field have come to define text mining in rather broad terms. For some, text mining centers on finding implicit information, such as associations between concepts, by analyzing large amounts of text. For others it pivots on extraction of explicit, not implicit, information from texts, such as named entities mentions or relations explicitly such as “A leads to B.” The task of identifying sentences with co-occurrences of a drug and a gene entity (for posterior manual curation into a database) is an example of the latter definition of text mining, which revolves around finding explicit information. Still, there are those who define text mining in the most stringent form: finding associations between a specific gene and a specific drug(s) based on clear-cut statistical analysis. No matter what view one subscribes to, text mining tools and methods are utilized, nonetheless, to significantly reduce human effort to build information systems and to automate the information retrieval and extraction process.

In particular, text mining aids in the search for information by using patterns for which the values of the elements are not exactly known in advance. In short, such tools are used to automate information retrieval and extraction systems, and by so doing, they help researchers to a large extent in dealing with the persistent problem of information overload. All in all, biomedical text mining “holds the promise of, and in some cases delivers, a reduction in cost and an acceleration of discovery, providing timely access to needed facts, as well as explicit and implicit associations among facts” (Simpson & Demner-Fushman 2012, p. 466). In this vein, biomedical text mining tools have been developed for the purpose of improving the efficiency and effectiveness of medical researchers, practitioners, and other health professionals so that they can deliver optimal health care. In the end, it is the patient who benefits from a more informed healthcare provider.

The field of text mining has witnessed a number of interesting applications. In Nahm and Mooney’s (2002) AAAI technical report on text mining they describe how a special framework for text mining, called DiscoTEX (Discovery from Text EXtraction), uses “a learned information extraction system to transform text into more structured data” so that it can be “mined for interesting relationships” (p. 60). In so doing, they define text mining as “the process of finding useful or interesting patterns, models, directions, trends or rules from *unstructured* text” (p. 61). In contrast to DiscoTEX, there are those applications that to try to infer higher-level associations or correlations between concepts.

Arrowsmith¹ and BITOLA² are examples of such text mining applications that work on this higher level of association. Similarly, both MEDIE³ and EvenMine⁴ are examples of systems that perform more fine-grained linguistic analysis.

In Feldman and Sanger's text mining handbook (2006) the authors show how text mining achieves its goal of extracting useful information from document collections "through the identification and exploration of interesting patterns." Though the authors show that "text mining derives much of its inspiration and direction from seminal research on data mining," they also emphasize that text mining is vastly different from data mining. This is so, because in text mining "the data sources are document collections" whereas in data mining the data sources are formal databases. As a result, in text mining, interesting patterns are found not among formalized database records" (as is the case with data mining), but rather "in the unstructured textual data in the documents in these collections" (p. 1).

Cohen and Hersh (2005) show that though text mining is concerned with unstructured text (as is likewise the case with natural language processing) it can, nevertheless, be "differentiated from ... natural language processing (NLP) in that NLP attempts to understand the meaning of text as a whole, while text mining and knowledge extraction concentrate on solving a specific problem in a specific domain identified *a priori* ..." The authors provide as an example the compilation of literature pertaining to migraine headache treatment, showing how the use of text mining "can aid database curators by selecting articles most likely to contain information of interest or potential new treatments for migraine [which] may be determined by looking for pharmacological substances that are associated with biological processes associated with migraine" (p. 58).

Current trends in biomedical text mining (Hakenberg et al. 2012; Gurulingappa et al. 2013; Zhao et al. 2014) include the extraction of information related to the recognition of chemical compound and drug mentions or drug dosage and symptoms. They also include extraction of drug-induced adverse effects, text mining of pathways and enzymatic reactions, and ranking of cancer-related mutations that cluster in particular regions of the protein sequence.

In this chapter, we explore how text mining is used to perform biomedical knowledge extraction. By describing its main phases, we show how text mining can be used to obtain relevant information from vast online databases of health science literature and patients' electronic health records. In so doing, we describe

1 http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

2 <http://ibmi3.mf.uni-lj.si/bitola/>

3 <http://www.nactem.ac.uk/medie/>

4 <http://www.nactem.ac.uk/EventMine/>

the workings of the four phases of biomedical knowledge extraction using text mining (text gathering, text preprocessing, text analysis, and presentation) entailed in retrieval of the sought information with a high accuracy rate. The chapter also includes an in depth analysis of the differences between clinical text found in electronic health records and biomedical text found in online journals, books, and conference papers, as well as a presentation of various text mining tools that have been developed in both university and commercial settings.

1.2 Background

1.2.1 Clinical and biomedical text

In general, clinical text is written by clinicians in the clinical setting. This text describes patients in terms of their demographics, medical pathologies, personal, social, and medical histories and the medical findings made during interviews, laboratory workup, imaging and scans, or the medical or surgical procedures that are preformed to address the underlying medical problem (Meystre et al. 2008). Here is an example of what clinical text may look like: “a sixty five year old Caucasian female with acute pancreatitis with history of gall stones ... patient complains of severe weight loss and abdominal pain ... blood test shows increase in blood serum amylase and lipase ... abdominal ultrasound shows enlarged bile duct ... ERCP (endoscopic retrograde cholangiopancreatography) scheduled for patient next week for removal of stones from bile duct ... patient to be placed on low fat diet ...” (Though in actual clinical notes, abbreviations and symbols, such as those that indicate the patient’s gender, are often used, we chose to omit such shorthand text for the purpose of giving a clear example here.)

As this example shows, clinical text describes a sequence of events and narratives, with the goal in mind of producing as precise and comprehensive an explanation as possible when describing the health status of a patient. This type of expressive description found in the clinical narrative understandably inheres a fair amount of ambiguity and personal differences in both vocabulary and style (Lovis et al. 2000; Suominen 2009). The main purpose of clinical text is to serve as a summary or “handover note” of patient care (documentation relating to the transfer of responsibility of the patient to another care provider either within the same healthcare setting or at another health facility), but it can also be used for legal requirements, care continuity, reimbursement, case management and research. Clinical text covers every phase of care, and depending on the purpose, the documents may differ in style, lengthiness, conformity to grammatical rules

and so on. As such, documents describing lab results and medical examinations are very different from those that describe patient care outcome in both the long run and short run.

There are other variations of clinical text as well. That is, clinical text may be entered either in *real time* or in retrospect, as a summary. In addition, clinical text may be entered at the patient's bedside or elsewhere (Thoroddsen et al. 2009). Clinical text contrasts with biomedical text, which is the kind of text that appears in books, articles, literature abstracts, posters, and so forth (Meystre et al. 2008). This is the kind of text that appears in MEDLINE/PubMed resources. Although both types of text do have some similarities, in that the heavy use of domain-specific terminology and the frequent inclusion of acronyms and polysemic words are found in both mediums, there are several features that make clinical text different from biomedical text. It is these differences that make clinical text especially challenging to NLP. Here are some of the reasons:

- Some clinical texts do not conform to the rules of grammar, are short, and are composed of telegraphic phrases;
- Clinical narratives are full of abbreviations, acronyms, and other shorthand phrases. Also, these shorthand lexical units are often overloaded, i.e., the same set of letters has multiple interpretations (Liu, Lussier & Friedman 2001);
- Misspellings are frequent in clinical text, as it is often produced without any spelling support;
- Clinical narratives often contain pasted sets of laboratory values or vital signs with embedded non-text strings, complicating otherwise straightforward NLP tasks like sentence splitting; and
- Templates and pseudo tables are often composed in plain text that are made to look tabular by the use of white space or lists.

Information search from this type of narrative text is difficult and time consuming. Standardization and structuring have been proposed as possible solutions. However, such solutions are not free of problems. For example, converting narratives to numerical and structured data is laborious and easily leads to differences and errors in coding. Moreover, if these tasks are performed manually, which is currently the most common approach, text ambiguity and personal differences may cause inconsistencies (Suominen 2009). Also, converting narratives into structured data may lead to significant information losses, as it limits the expressive power of free-text (Lovis et al. 2000; Walsh 2004).

1.2.2 Information retrieval

The term “Information Retrieval” was coined in 1952; a decade later this term came to be popularly used in the research community (Van Rijsbergen 1979) and has continued to date. When the first automated information retrieval systems were actually introduced during the 1960s, the field of information retrieval (IR) was born. Information retrieval can be defined as the art and science of searching for information in large collections of documents; and, likewise, searching for text, sound, or images within those documents themselves. In addition, the search for metadata about documents is also part of information retrieval. According to Manning, Raghavan and Schutze (2008) “Information Retrieval (IR) is finding documents of an unstructured nature that satisfies an information need from within large collections (usually stored on computers).” As such, the field of information retrieval (IR) is the study of techniques for organizing and retrieving unstructured text stored on the computer. However, working with unstructured text, such as web pages, text documents, office documents, presentations and emails, can be quite difficult. That is, since unstructured text does not have a data model, it cannot be easily processed by a machine. *Structured* data, on the other hand, is either, in general, annotated or contained in databases (e.g., library catalogues and phone numbers), whereas unstructured data is not. (See Appendix “A” for list of open-sourced structured databases.)

Singhal (2001) opined that since the quantity of electronic information has increased dramatically with the widespread adoption of World Wide Web during the 1990s, information retrieval has become a sphere of great interest. Similarly, he saw the research and growth in this area as a natural consequence of the increasing interest in information retrieval.⁵

⁵ To support research within the IR community, a special program was erected by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense back in 1992. The program was called The Text Retrieval Conference (TREC), which is part of the TIPSTER Text program. TREC consist of an ongoing series of workshops focusing on a list of different IR research areas or tracks. Its purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies and to increase the speed of lab-to-product transfer of technology. The TREC test collections and evaluation software are available to the retrieval research community writ large, so organizations can evaluate their own retrieval systems at any time. TREC has successfully met its dual goals of improving the state-of-the-art in information retrieval and in facilitating technology transfer. Retrieval-system effectiveness approximately doubled in the first six years of TREC. In TREC-2003 there was a retrieval track dedicated to Genomics, and in 2004 this track was centered on tagging genes and proteins in relevant documents: <http://trec.nist.gov/>.

1.2.2.1 Information retrieval process

Information retrieval is used to locate specific items in a set of natural-language documents, such as finding specific gene-related information from the biomedical literature. IR systems provide a way for a user to enter a *query*, using keywords, wherein the system will return the documents considered relevant to the query from the document collection. To do so, Herrera-Viedma (2001) explains, “both documents and user queries must be formally represented in a consistent way so that IRS [Information Retrieval System] can satisfactorily develop the retrieval activity.” IR is achieved by scanning the collection for matched terms when a search is performed. The author shows that, basically, three components are involved in the information retrieval process:

1. *A Database*: which stores the documents and the representation of their information contents (index terms). It is built using tools for extracting index terms and for representing the documents.
2. *A Query Subsystem*: which allows users to formulate their queries by means of a query language.
3. *An Evaluation Subsystem*: which evaluates the documents for a user query. It presents an inference procedure that establishes a relationship between the user request and the documents in the database to determine the relevance of each document to the user query (p. 460).

The author points out that to help overcome the “lack of flexibility and precision for representing document contents, for describing user queries and for characterizing the relevance of the documents retrieved for a given user query” weights are incorporated at these three levels of information representation. Namely, at the document representation level in which a database is built, “by computing weights of index terms, the system specifies to what extent a document matches the concept expressed by the index terms”; at the query representation level “by attaching weights in a query” which allows the user to “provide a more precise description of his or her information needs or desired documents”; and at the evaluation representation level “by assigning weights to characterize the relationship between user queries and document representation” so that the evaluation subsystem can provide a means, known as the retrieval status value (RSV) of a document “to discriminate the documents retrieved by relevance judgments” (pp. 460–461).

In fact, a number of researchers in the field of informational retrieval have been encouraged to devise ways of making the entire information retrieval process more efficient. Some have, for example, embarked on various

ways of streamlining the index size of the IR system. Gonzalez (2008) showed how the index system, also known as the inverted file (IF) that “serves as the data structure in charge of storing the information handled in the retrieval process” can be compressed using “document reordering and static index pruning.” The author shows how this new approach differs from the traditional “static compression schemes” though they are deemed “complementary to them,” and that regardless of the approach used they all “have one thing in common: they make use of some of the properties inherently related to document collection.”

1.2.3 Information extraction

Information extraction (IE) systems analyze *unstructured* text in order to extract information about pre-specified types of events, entities or relationships, such as the relationship between disease and genes or disease and food items. In other words, information extraction is all about deriving structured information from unstructured text. This differs from information retrieval (IR), described above, in that the purpose of IE is to *add* value and insight to the data whereas IR simply locates information in the same form(s) that it is stored without supplying any additional analytical insight about correlations, co-morbidity, or any other co-occurrence.

In addition, IE may be seen as a subtask of text mining, since the latter is a vast area that includes document classification, document clustering, building ontologies and other tasks, whereas IE is primarily concerned with crawling, parsing, and indexing documents so as to extract useful information from the data. In recent years, however, IE has distinguished itself from text mining as multimedia document processing, involving automatic annotation and content extraction from images, audio and video clips, has become more widely used. In fact, radiologists have come to depend on information extraction from medical images, using automatic image annotation systems in some of the more novel and creative ways.

1.2.4 Challenges to biomedical information extraction systems

Biomedical information extraction can build a database with the information on a given relationship or event drawn from a variety of sources such as online medical news, biomedical literature, or electronic health records. Since the

documents are *unstructured* and expressed in a natural language format, it is very difficult for a computer to understand and analyze them. Yet, scientists and clinicians need to keep up-to-date with all of the new discoveries and theories presented in the biomedical literature, and they must, likewise, make efficient use of this ever-expanding reservoir of biomedical information. Undoubtedly, there is a significant degree of information overload.

Not surprisingly, information overload places a heavy burden on biomedical information extraction systems to perform efficiently. However, biomedical IE systems face yet another problem, one that is undoubtedly *sui generis* to the biomedical domain. Ramampiaro (2010) describes how medical terms often cross over to vernacular usage, thereby causing false positives that artificially boost ranking scores. The duality of meaning ascribed to words, which can be found in both the vernacular or, alternatively, in biomedical literature and in clinical documents, constitutes a persistent problem associated with biomedical IE. Krauthammer and Nenadic (2004) point out that this duality of usage presents one of the biggest challenges to biomedical extraction in that “biomedical information typically contains large amounts of domain-specific terminology with *high ambiguity*” (emphasis supplied). This makes indexing particularly difficult.

For example, *heart* means the hollow muscular organ located behind the sternum and between the lungs in the medical context, but in the vernacular English language, it may be used to convey a different meaning, as in “the child won everyone’s *heart*.” Such linguistic ambiguities may create serious problems with how to rank the documents at hand. Finding the occurrence of the word “heart” many times in an online news article, for example, may give a speciously high ranking to the document if indeed the word “heart” had been used in a vernacular rather than in a biomedical context.

Furthermore, the need to learn and derive new knowledge also remains a challenge for biomedical information extraction systems. For all these reasons, there remains a growing need for the development of effective tools to meet these challenges and obstacles head-on so as to enable researchers and practitioners (and lay members who may need to research certain health issues) to access and extract useful information from the biomedical literature. It is understandable that this will require better machine learning tools that can perform heuristic discoveries so as to learn new relationships between entities and events that are not previously stored in the system.

In addition, the rapid increase in the sheer volume of biomedical literature necessitates the design of information extraction tools similar to the “open discovery” algorithm introduced by Srinivasan and Libbus (2004), which they used

to “uncover information that could form the basis of new hypotheses.” Or, the MedMeSH Summarizer System described by Kankar et al. (2002) to help streamline the process of cross-referencing “experimental and analytical results with previously known biological facts, theories, and results.” This is much needed given the breadth of biomedical databases, which can ordinarily make “the task of cross-referencing very lengthy, tedious, and daunting.”

In sum, it is these special requirements of the biomedical domain that call for a new set of text mining tools, since the tools used for other domains have not proven entirely successful when applied to the biomedical sciences.

1.2.5 Applications of biomedical information extraction tools

Information extraction tools are used across various domains such as security, online media, marketing applications (Coussement & Poel 2008), and web mining (Zanasi 2009). Biomedical information extraction tools are used to perform a variety of functions. Text mining applications in biomedical area are diverse and they include:

1. The identification of chemical compounds: identifying their structures and the relations between them; and identifying drugs in which the particular compound is used, along with their respective side effects and toxicity (Vazquez et al. 2011);
2. Disease research such as cancer: several applications were developed to provide easy access to the most recent developments in cancer research (Zhu et al. 2013);
3. Genetics: gathering the most recent information about complex processes involving genes, proteins and phenotypes (Jensen, Jensen & Brunak 2012; Rebholz-Schuhmann, Oellrich & Hoehndorf 2012);
4. Extracting gene-based patterns using natural language processing techniques to extract the rhetoric information (the intention to be conveyed to the reader by the author(s) of the paper) contained in technical abstracts (Atkinson, Ferreira & Aravena 2004);
5. Indexing Medline documents (Kankar et al. 2002);
6. Finding the relationship between curcumin longa (a dietary substance) and retinal diseases (Srinivasan, Bisharah & Sehgal 2004);
7. Developing an expert system to perform medical diagnosis from clinical patient records and patient histories (Moumtzoglou & Kastania 2011); and
8. Finding risk factors of a disease (Imambi & Sudha 2010).

1.3 Biomedical knowledge extraction using text mining

The main phases, as shown in Fig. 1.1, of biomedical knowledge extraction using text mining are: (1) Unstructured text gathering and preprocessing; (2) Extraction of features and semantic information (including information extraction and creation of semantic metadata) to produce annotated texts; (3) Analysis of the annotated texts (using data mining, semantic search and knowledge discovery); and (4) Presentation. Each phase will be discussed in turn.

Typical text mining applications include the following: identification of facts in specialized (domain-based) literature, discovery of implicit and unknown facts, document summarization, and entity-relation modeling (i.e., learning relations between named entities). Applications usually scan sets of document to identify relevant information. The relevant information can be identified by either modelling the document set, using one or more classification schemes, or populating a database (adding information to a database or adding fields to a database in order to be able to fill it with information) or search index with the information that is extracted.

Some important subtasks are:

- Information retrieval or identification of a corpus, a preparatory step for collecting or identifying a set of textual materials (that either appear on the Web or are held in a file system, database, or content management system) for analysis.
- Named entity recognition is the use of gazetteers or statistical techniques to identify named text features: diseases, drugs, anatomical structures, dysfunctions, lab procedures, certain abbreviations, and so on. Disambiguation by using contextual clues that may be required in order to decide whether, for instance, “block” refers to a specific medical condition such as *intraventricular* block or *heart* block, or some other entity for that matter.
- Natural language processing (which are considered complex tasks that can take more time to complete), such as part of speech tagging, syntactic parsing, and other types of linguistic analysis. These tasks are performed less



Fig. 1.1: Main phases of biomedical knowledge extraction using text mining.

frequently, in part, as they require a longer processing time. Machine learning approaches usually include these tasks to generate features to be analyzed in the learning process and to support decision in runtime. Features can be at the token level, as lemmas and part of speech tags, or at the sentence level using syntactic parsing.

1.3.1 Unstructured text gathering and preprocessing

1.3.1.1 Text gathering

The text-gathering phase provides an “interface” to collect the raw documents from online sources, such as online journals, books, and conference papers and from electronic health records compiled at major teaching hospitals and at local community medical facilities. Biomedical information is, thus, made available through such online literary databases and health records, as well from the web in general. One such interface for published materials is PubMed, whose largest component is MEDLINE, which serves as a freely accessible online database of biomedical journal citations and abstracts created by the U.S. National Library of Medicine.

As of 2014, Medline includes citations from over 5600 scholarly journals published in more than 80 countries around the world. PubMed comprises more than 22 million references that include the entire MEDLINE database and other types of citations, such as in-process citations, which provide records for articles before those records go through quality control and are indexed; citations to articles that are out-of-scope from certain MEDLINE journals; citations that precede the date that a journal was selected for MEDLINE indexing; and other works such as chapters and books that are likewise outside the purview of MEDLINE.⁶ This repository of scientific literature provides a vast amount of text data that has helped researchers to implement their classification algorithms (Imambi & Sudha 2011).

The electronic health record (EHR) constitutes another major source of digital web-based data, primarily existing as part of the hospital’s own collection of private computer networks (*an intranet*) rather than as part of the World Wide Web. Yet, this source of data can serve a goldmine of valuable clinical and demographic information on patient care. Such records contain a large repository of Patient Notes that describe the patient’s medical history and treatments, plans for follow-up treatment after the patient is discharged from the hospital, the test results and lab reports of the patient during in-hospital care, and the many other aspects of

⁶ https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html

patient care that had not been captured in the structured part of the EHR. The information in the notes can be found in the form of descriptive and semi-structured format. These data can be mined for genomic research purposes as well. In fact, Denny (2012) showed that “when linked to biological data such as DNA and tissue biorepositories, EHRs can become a powerful tool for genomic analysis.”

1.3.1.2 Text preprocessing

The document set obtained is prepared for processing. First, the document text is tokenized. Tokenization is the division of a text into meaningful units called “tokens.” A token is a group of characters that is categorized according to a set of rules. For instance, NUMBER, COMMA and DOT are examples of token categories. It is an important task since all the following tasks will be based on tokens resulting from this process. Thus, several tokenization solutions were developed for several domains and languages. For instance, OpenNLP⁷ has models for biomedical documents in English and Portuguese, and SPECIALIST NLP (Browne, McCray & Srinivasan 2000) supports English biomedical text. This process is also referred to as “feature generation.”

Next, some words are removed. These words are called *stop words*. They consist of words that are frequently used, such as “it,” “are,” and “is.” Though such words are quite common, since they are not useful in the classification of documents they are summarily removed (National Center for Biotechnology Information 2010).

Since in most cases morphological variants of words have similar semantic interpretations, which can be considered as equivalent, words are stemmed as part of preprocessing. Word stemming reduces inflected and derived word forms to their root or stem, mapping related words to the same stem, for example, the words “retrieval,” “retrieve,” and “retrieving” become *retrie* when stemmed.

1.3.2 Extraction of features and semantic information

This next phase usually starts with named entity recognition (NER), which aims to detect specific terms that represent relevant entities such as genes, proteins, diseases, and drugs. There still exist important challenges in named entity recognition that derive from the fact that there are different ways of referring to the same phenomena. For instance, “epilepsy” and “falling sickness” refer to the same disease: a central nervous system disorder characterized by the loss of consciousness (Zhu et al. 2013).

⁷ <http://opennlp.apache.org/>

The natural language text of biomedicine, found in articles, books, reports, and other unstructured sources, present several challenges that can make the application of information extraction and retrieval techniques even harder. The main challenge is related to terminology, and is a result of the complexity of the terms used in biomedical entities and processes (Zhou et al. 2004; Ananiadou & McNaught 2006):

- Non-standardized naming convention: an entity name could be found in various spelling forms (e.g., “N-acetylcysteine,” “N-acetyl-cysteine,” and “NAcetylCysteine”);
- Ambiguous names: a same name could be related with more than one entity, depending on the text context;
- Abbreviations: biomedical abbreviations are frequently used (e.g., “TCF” may refer to “T cell factor” or to “Tissue Culture Fluid”);
- Descriptive naming convention: many entity names are descriptive, which makes its recognition a complex task (e.g., “normal thymic epithelial cells”);
- Conjunction and disjunction: two or more entity names sharing one head noun (e.g., “91 and 84 kDa proteins” refers to “91 kDa protein” and “84 kDa protein”);
- Nested names: one name may occur within a longer name, as well as occur independently (e.g., “T cell” is nested within “nuclear factor of activated T cells family protein”)
- Names of newly discovered entities: there is an overwhelming growth rate and constant discovery of novel biomedical entities, which takes time to register in curated nomenclatures.

In general, there have been several approaches to NER in the clinical and biomedical literature. These can be roughly divided into the following four groups: (1) Dictionary-based approaches that try to find names of the well-known nomenclatures in texts; (2) Rule-based approaches that manually or automatically construct rules and patterns to directly match them to candidate named entities in the texts; (3) Machine learning approaches that employ machine learning techniques, such as Hidden Markov Models and Support Vector Machines, to develop models for NER; and (4) Hybrid approaches that merge two or more of the above approaches, mostly in a sequential way, to deal with different aspects of NER.

1.3.3 Analysis of annotated texts

In this next phase, various text mining techniques can be applied to the preprocessed data. Frequent tasks associated with this phase are the following:

Relation extraction: After having identified named entities, several information extraction tasks in the biomedical domain involve determination of

relationships among those entities. The goal of the relation extraction task is to identify occurrences of particular types of relationships between pairs of entities. Although common entity classes, such as genes or drugs, are in general quite specific, relations may be broad, including any type of biomedical association. Alternatively, such relations may be very specific, for example, by characterizing only gene regulatory associations (Simpson & Demner-Fushman 2012). Relation extraction approaches have shown an evolution from simple systems that rely solely on co-occurrence statistics to complex systems utilizing syntactic analysis and dependency parsing.

Event detection: Recently, there has been a shift in biomedical information extraction from recognizing binary relations to the more ambitious task of identifying complex, nested event structures. Events are typically characterized by verbs or nominalized verbs. For example, in the sentence “glnAP2 may be activated by NifA,” the verb activated specifies the event, and glnAP2 and NifA are the event’s arguments. Unlike the case of simple binary relations, both concept labels and semantic roles are assigned to an event and its arguments. In this example, the verb activated indicates a positive regulation type event, which expects a protein (NifA) to act as the event’s cause and a gene (glnAP2) to act as the event’s theme (Ananiadou et al. 2010).

Semantic search and inference: Search in large collections of documents, as those in biomedical and health domains, presents a series of challenges. A highly relevant one is vocabulary mismatch because it can severely decrease the performance of keyword-based search. This can happen when a user’s query contains little or no shared terms with relevant documents for that query. For example, when querying “lung cancer treatment,” documents using specialized terms such as “lung excision” or “chemotherapy” may receive a low rank or even be left out of the result set altogether. Vocabulary mismatch is dealt with by using techniques such as query term expansion and inference (Liu & Chu 2007; Koopman et al. 2011).

Text summarization: Medical information is often fragmented, existing in a wide range of locations and formats. This fragmentation makes the creation of an optimal clinical summary more challenging (Febowitz et al. 2011). The availability of a great amount of clinical information that can be accessed rapidly increases the risk of inefficacy due to information overload (Hall & Walton 2004). This problem is likely to increase over time with the sharing of patient data more broadly. This makes clinical text summarization an important task. It can be divided into three interrelated categories: source-oriented, time-oriented and concept-oriented views (Febowitz et al. 2011).

Text clustering: The objective is to organize text in a small number of meaningful clusters of the same type or class. Classes are usually obtained

from the set of relevant and frequent words of the text, and thus the number of classes that will be assigned is not known beforehand. Text clustering finds applicability for a number of tasks, such as document organization and browsing, corpus summarization, and document classification (Simpson & Demner-Fushman 2012).

Automated text categorization: Is the process of assigning unseen documents to user-defined categories. An important goal in biomedical text mining is automatic classification of electronic documents. Computer programs scan text in a document and generate a model that assigns the document to one or more pre-specified topics/categories using classification techniques. Those categories are usually organized in taxonomies (Fang, Parthasarathy & Schwartz 2001). Text classification, adopted as an example, is the subject of next section.

1.3.3.1 Algorithms for text classification

Several approaches have been proposed. Text classification is based on the supervised learning model. In this learning the total documents are divided into two parts. One part is called “training data” and the other part is called “test data.” A model or classifier is generated with training data. Once a classifier is created, it is applied to test the dataset in order to calculate the accuracy of the classifier. The frequently used text classification algorithms are Naïve Bayesian, k-NN, Decision Trees, and SVM.

Naive Bayesian (NB) algorithm

Naïve Bayesian (NB) algorithm has been generally used for text classification. This algorithm is based on Bayes’ theorem and is used to predict the probability of categories for a given document. The classifier predicts posterior probability of documents for each category and assigns the category which has highest posterior probability. Naive Bayesian classifier assumes that the effect of the probability of the term on a given category is independent of the probability of the other terms in the same category (Zhang, Chen & Xiong 2007; Yuan 2010).

There are two versions of the NB algorithm. One is the multi-variate Bernoulli event model that only takes into account the presence or absence of a particular term so that it doesn’t capture the number of occurrences of each word. The other model is the multinomial model that captures the word frequency information in documents. Li and Jain (1998) showed that Naïve Bayesian classifier does not provide efficient classification with smaller training data sets. If the training set is limited in size, then there may be a chance that the term frequency of some of

words will become zero and, at the same time, the probability of the word in a given category also becomes zero.

k-Nearest Neighbor algorithm (k-NN)

k-NN classifier is an instance-based learning algorithm that is based on a distance function for pairs of observations, such as the Euclidean distance or cosine. In this classification process, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. k is a positive integer, typically small. If $k = 1$, then the object is simply assigned to the class of its nearest neighbor. In binary classification problems, it is helpful to choose k to be an odd number as this avoids tied votes.

“One of the advantages of k-NN is that it is well suited for multi-modal classes as its classification decision is based on a small neighborhood of similar objects. So, even if the target class is multi-modal it can still lead to good accuracy.”⁸ The drawback of k-NN is that it uses all features in the documents to compare them. It affects the similarity measure and consequently the efficiency of classification. The traditional k-NN text classification algorithmic limitations are: calculation complexity mainly due to the usage of all the training samples for classification; dependency on the training set; and equal weighting of all samples. To overcome these challenges researchers developed variations of k-NN algorithms.

Decision trees

Decision trees are one of the most widely used inductive learning methods. Decision tree algorithms are suitable for document classification because of their robustness to noisy data. Two widely known algorithms for building decision trees are classification and regression trees. ID3 and its successor C4.5 (Quinlan 1993) and booster version of C 4.5 (Quinlan 1998) are famous for classification. It is a top-down approach which recursively constructs a decision tree classifier. At each level of the tree, ID3 selects the attribute that has the highest *information gain*. “ID3 is a supervised machine learning algorithm that automatically derives a decision tree from a set of training instances once each instance is tagged with its correct classification. A fully trained decision tree can then be used to classify previously unseen instances from a test set” (Lehnert et al. 1995). The tree tries

⁸ <http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab1-Algorithms%20for%20Information%20Retrieval.%20Introduction.pdf>

to split the training data based on the values of the available features to produce a good generalization. The node which has highest information gain is used to make a split. Each leaf node represents a class label. The given document is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that document. The leaf node reached is considered as the class label for that document. Decision tree algorithms are suitable for both binary and multiclass classification.

Support vector machines

Support vector machine (SVM) is a popular technique for classification. In recent years, the SVM has become an effective tool for pattern recognition, machine learning, and data mining because of its high generalization performance. The goal of SVM is to produce a model that predicts target value of data instances in the testing set, which are only given the attributes. Support vector machines (SVM) is a new technique for data mining, which has received increasing popularity in the machine learning and statistics community. SVM has been introduced by Vapnik (1995) for solving pattern recognition and nonlinear function estimation problems. SVM has become the tool of choice for the fundamental classification problem of machine learning and data mining. “Unlike traditional methods which minimize the empirical training error, SVM aims at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyperplane and the data. This can be regarded as an approximate implementation of the structure risk minimization principle” (Wang, Chen & Chen 2004, p. 512).

Support vector machines are among the most robust and successful classification algorithms. They are based upon the idea of maximizing the margin i.e., maximizing the minimum distance from the separating hyperplane to the nearest example. The basic SVM supports only binary classification, but several extensions of these algorithms can deal with multiclass classification as well (Bredensteiner & Bennett 1999). SVM is frequently used in the medical domain. For example, it is used to generate a decision support system for heart disease classification (Bhatia, Prakash & Pillai 2008).

1.3.3.2 Classification evaluation measures

The evaluation is essential for understanding the quality of the learning model, for tuning the parameters in the iterative process of classification, and for selecting the best model. There are several measures for evaluating models such as complexity, computational cost, computational time, mean absolute error, sensitivity, specificity, and accuracy.

Confusion matrix

A classification model classifies each instance into one of the classes. The confusion matrix shows how the predictions are made by the model. The rows correspond to the class labels in the data set. The columns show the predictions made by the model. The value of each element in the matrix is the number of predictions made with the class corresponding to the column. Thus, the diagonal elements show the number of correct classifications made for each class, and the off-diagonal elements show the errors made.

There are four possible classifications for each instance: i.e., true positive, true negative, false positive, and false negative. This is represented in matrix form and is called confusion matrix. If the accuracy of the classification model is 100% then all predictions are correct, which means that false positives and false negatives have a value of zero. The below Tab. 1.1 shows how the results are tabulated in a confusion matrix.

Mean absolute error

The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to eventual outcomes. The mean absolute error is given by

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |f_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i|$$

The mean absolute error is an average of the absolute error $e_i = |f_i - y_i|$, where f_i is the prediction and y_i is the true value.

Kappa statistics

Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance

Tab. 1.1: Confusion matrix.

		Observed	
		True	False
Predicted	True	True Positive rate (tp)	False Positive rate (fp)
	False	False Negative rate (fn)	True Negative rate (tn)

away from the observed agreement and dividing by the maximum possible agreement:

$$K = \frac{P_o - P_c}{1 - P_c}$$

where P_o is the proportion of observed agreement and P_c is the proportion of agreements expected by chance. A value greater than “0” means the classifier is doing better than chance.

Accuracy

Accuracy is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision

Precision is a measure of the accuracy provided that a specific class has been predicted. Precision is the probability that a retrieved document is relevant. From the confusion matrix it is calculated by:

$$Precision = \frac{tp}{tp + fp}$$

where tp and fp are the numbers of true positive and false positive predictions for the considered class. Precision is 1 when fp is 0, which indicates there were no spurious results.

Recall

Recall is the probability that a relevant document is retrieved in a search. Recall is also referred to as the true positive rate or sensitivity and is given by:

$$Recall = \frac{tp}{tp + fn}$$

Recall becomes 1 when fn is 0, and it indicates that 100% of the tp were discovered.

F-measure

The F-measure is the harmonic mean of precision and recall. It is calculated by using the formula:

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The behavior of the performance measures is the function of the decision threshold for classification. When decision threshold increases, the recall will increase and precision will decrease.

1.3.4 Presentation

In this last phase, the result of classification is represented in graphical format, so that even non-technical people can also easily interpret the result. There are several presentation tools available. These tools are also called *data visualization tools*. Some of them are Plotly,⁹ IBMMany Eyes,¹⁰ Grapheur,¹¹ Visumap¹², etc. These tools are not only used to represent the relationships and co-relations, but they are also used to represent patterns of data.

1.4 Text mining tools

Text mining tools help in discovering structure and patterns in unstructured data – usually text. These tools are available from many commercial and open source companies. Some relevant general-purpose tools are:

SAS Text miner: This tool extracts knowledge from unstructured data with text mining software. It provides interactive GUIs which makes it easy to identify relevance, modify algorithms, document assignments, and group materials into meaningful aggregations. This makes it easy for the user to guide machine-learning results with human insights. It extends text mining efforts beyond basic start-and-stop lists by using custom entities and term-trend discovery to refine automatically generated rules.¹³

9 <https://plot.ly>

10 <http://services.alphaworks.ibm.com/manyeyes/>

11 <http://www.grapheur.com/>

12 <http://www.visumap.net/>

13 <http://www.sas.com>

NetOwl Text Analytics: NetOwl offers a suite of best-of-breed text and entity analytics products. “NetOwl analyzes Big Data in the form of text data – news, email, web, social media, and any other text document that organizations would like to exploit as well as structured entity data about people, organizations, places, and things.”¹⁴ It provides tools to analyze an extremely large volume of data in a variety of forms and languages and offers advanced text analytics products to meet today’s Big Data challenges.

IBM Intelligent Miner: IBM Intelligent Miner for Text is a knowledge discovery software development toolkit. It contains tools for application programmers who want to build applications to extract key information from very large quantities of documents, e-mails, or Web pages stored online, often on the Internet or on intranets, without having to read them all. IBM Text Analysis Tools include a Language Identification tool, comprehensive Clustering tools, a Topic Categorization tool, a Summarization tool, and Feature Extraction tools. These tools identify document language, group conceptually related documents, classify documents by content, generate document summaries, and extract key elements of text.¹⁵

Weka: WEKA is an open-source machine learning tool. It was developed at the University of Waikato, New Zealand to implement data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, and association rules; it also includes visualization tools. The new machine learning schemas can also be developed with this package. WEKA is open-source software issued under General Public License.¹⁶

Adding to these general purpose tools, some specialized tools were developed for specific topics related to biomedical and health domains. Simpson and Demner-Fushman (2012) present a comprehensive review of recent works; an extensive list can be found in the Bio-NLP resources database.¹⁷ Some relevant systems are:

Becas: becas¹⁸ is a web application, API, and widget for biomedical concept identification that helps researchers, healthcare professionals, and

¹⁴ <http://www.netowl.com>

¹⁵ <http://www-01.ibm.com/common/ssi/cgi-bin>

¹⁶ <http://www.cs.waikato.ac.nz/ml/weka>

¹⁷ http://zope.bioinfo.cnio.es/bionlp_tools/get_all_bionlp_tools_out?SUBMIT#equal#Submit+Query

¹⁸ <http://bioinformatics.ua.pt/becas/#/>

developers in the identification of over 1,200,000 biomedical concepts in text and PubMed abstracts (Nunes et al. 2013). It provides annotations for isolated, nested, and intersected entities, and identifies concepts from multiple semantic groups. It has the ability to provide preferred names for concept identification and is able to enrich them with references to public knowledge resources.

KLEIO: enhances search facilities across the MEDLINE collection by identifying key entities within the text, such as gene names or proteins, and improves the querying method with unique identifiers by automatically including synonyms, spelling variants and, even, disambiguating acronyms (Nobata et al. 2008). It combines these features with the common features found in other interfaces to provide a solution to the growing problem of finding valuable information within the ever increasing volume of modern publications.¹⁹

PIE the search: *PIE* (Protein Interaction information Extraction) *the search* is a web service to extract protein-protein interaction relevant articles from MEDLINE (Kim et al. 2012). It accepts PubMed input formats to make available up-to-date protein-protein interaction information which cannot be found in manually curated databases. *PIE the search* is targeted at providing protein-protein interaction relevant articles for biologists, baseline system performance for bio-text mining researchers, and a compact PubMed-search environment for PubMed users.²⁰

MEDIE: is a framework for accurate, real time, retrieval of relational concepts from MEDLINE (Miyao et al. 2006). Prior to retrieval, a semantically annotated text base is prepared and stored in a structured database. The preparation of the text base includes applying natural language processing tools, including deep parsers and term recognizers. User requests are converted on the fly into patterns of these semantic annotations, and texts are retrieved by matching these patterns with the pre-computed semantic annotations. Real-time retrieval is possible because semantic annotations are computed in advance.²¹

MedInX: is a Medical Information eXtraction system tailored to process textual clinical discharge records, performing automatic and accurate mapping of free reports onto a structured representation (Ferreira, Teixeira & Cunha 2012). MedInX is designed to be used by health professionals, and by hospital administrators

¹⁹ <http://www.nactem.ac.uk/Kleio/>

²⁰ <http://www.ncbi.nlm.nih.gov/CBBresearch/Wilbur/IRET/PIE/>

²¹ <http://www.nactem.ac.uk/tsujii/medie/>

and managers, allowing a search of the contents of its automatically populated ontologies. (Further details on this system can be found in Chapter 3 of this book.)

NextBio: aggregates large quantities of genomic data for research and clinical applications. It contains the world's largest repository of curated and correlated public and private genomic data, including data from multiple public repositories of genomic studies and patient molecular profiles, up-to-date reference genomes, and clinical trial results (Kupersmidt et al. 2010). Several molecular data types from these resources are systematically processed, curated, and integrated into the data center based platform. This allows applying genomic data in novel and useful ways, both in the research laboratory and in the clinic.²²

The Neuroscience Information Framework: is a dynamic inventory of Web-based neuroscience resources: data, materials, and tools (Akil, Martone & Van Essen 2011). It helps in advancing neuroscience research by enabling discovery and access to public research data and tools worldwide through an open source networked environment. It offers the following: a search portal for researchers, students, or anyone looking for neuroscience information, tools, data, or materials; access to content normally not indexed by search engines; and tools for resource providers to make resources more discoverable, such as ontologies, data federation tools, and vocabulary services.²³

1.5 Summary

This chapter shows how biomedical information is successfully retrieved by using text mining techniques. The sources of biomedical information, found in both clinical narratives and biomedical literature, and the available tools for text mining are described in this chapter, which highlights various text mining techniques and evaluation measures. Future work, however, requires an interdisciplinary approach to text mining of biomedical information. Such coordinated efforts of biologists and clinicians, medical researchers and epidemiologists, computer scientists and computational linguists, library scientists and statisticians, and others are imperative to exploit the full scientific potential of biomedical text mining. The field has promise but much more effort must be made in choosing tasks and evaluating results based on real-world requirements and needs. In the end it is the patient population and the public writ large who will reap the full benefits of the application of text mining tools that successfully perform biomedical knowledge extraction.

²² <http://www.nextbio.com/b/nextbioCorp.nb>

²³ <http://www.neuinfo.org/>

Appendix “A”

Open-sourced Structured Databases

- Diseases Database:²⁴ It provides Cross-referenced database of clinical medicine and it links to topic categorical pages from other websites.
- DynaMed:²⁵ A medical information database with over 2000 diseases.
- General Practice Notebook:²⁶ Database of clinical medicine with a search facility.
- ICD-9 Data:²⁷ Offers drillable dataset of ICD-9-CM medical diagnosis codes.
- ICD-9 Search:²⁸ Search ICD-9 for medical diagnosis, codes, and procedures. Find related diseases, treatments and related news.
- ICD-9-CM Online:²⁹ Searchable database of disease classification.
- IndMED:³⁰ Indian Biomedical Journals Database: Bibliographic aggregation of peer-reviewed biomedical journals.
- OpenMED:³¹ An international open-access archive of scientific and technical documents for Medical and Allied Sciences.
- AIDSILIN database: It provides the literature on AIDS and HIV back to 1980.
- AMED Database:³² This database covers a range of complementary and alternative medicine including homeopathy, chiropractic, and acupuncture and so on.
- Bandolier:³³ Award-winning summary journal with searchable index produced by Andrew Moore and colleagues in Oxford, UK.
- Cochrane database.³⁴
- English National Board Health Care Database:³⁵ A database of journal references primary of interest to nurses, midwives and health visitors.

24 <http://www.diseasesdatabase.com>

25 <https://dynamed.ebscohost.com>

26 <http://www.gpnotebook.co.uk/homepage.cfm>

27 <http://www.icd9data.com>

28 <http://www.lumrix.net/icd-9.php>

29 <http://icd9cm.chrisendres.com>

30 <http://medind.nic.in/imvw>

31 <http://openmed.nic.in/>

32 <http://www.silverplatter.com>

33 <http://www.medicine.ox.ac.uk/bandolier/>

34 <http://www.mcmaster.ca/Cochrane/Cochrane/revabstr/abidx.htm>

35 <http://www.enb.org.uk/hcd.htm>

- POPLINE database:³⁶ The world's largest online bibliographic database on population, family planning, and related health issues. It is also available in CD-ROM which is free of charge to developing countries.
- STRIDE Clinical Data Warehouse³⁷ is the source of historical clinical data from both hospitals for research purposes.

References

- Akil, H., Martone, M. E. & Van Essen, D. C. (2011) 'Challenges and opportunities in mining neuroscience data', *Science*, 331:708–712.
- Ananiadou, S. & McNaught, J. (2006) 'Text mining for biology and biomedicine', *Comput Ling*, 135–140.
- Ananiadou, S., Pyysalo, S., Tsujii, J. & Kell, D. B. (2010) 'Event extraction for systems biology by text mining the literature', *Trends Biotech*, 28:381–390.
- Atkinson, J., Ferreira, A. & Aravena, E. (2004) 'Discovering implicit intention-level knowledge from natural-language texts', *Knowl-Based Syst*, 22:502–508.
- Bhatia, S., Prakash, P. & Pillai, G. N. (2008) 'SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features', In *Proceedings of the World Congress on Engineering and Computer Science*, pp. 34–38.
- Bredensteiner, E. J. & Bennett, K. P. (1999) 'Multi category classification by support vector machines', In *Computational Optimization*, Heidelberg, Germany: Springer. pp. 53–79.
- Browne, A. C., McCray, A. T. & Srinivasan, S. (2000) 'The specialist lexicon', *Natl Libr Med Tech Rep*, 18–21.
- Cohen, A. M. & Hersh, W. R. (2005) 'A survey of current work in biomedical text mining', *Briefings in Bioinformatics*, 6(1):57–71.
- Coussement, K. & Poel, V. D. (2008) 'Integrating the voice of customers through call center e-mails into a decision support system for churn prediction', *Inform Manage*, 45(3):164–174.
- Denny, J. C. (2012) 'Mining electronic health records in the genomics era', *PLoS Comput Biol*, 8(12).
- Fang, Y. C., Parthasarathy, S. & Schwartz, F. (2001) 'Using clustering to boost text classification', In *ICDM Workshop on Text Mining (TextDM'01)*.
- Febowitz, J. C., Wright, A., Singh, H., Samal, L. & Sittig, D. F. (2011) 'Summarization of clinical information: A conceptual model', *J Biomed Inform*, 44:688–699.
- Feldman, R. & Sanger, J. (2006) *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge: Cambridge University Press.
- Ferreira, L., Teixeira, A. & Cunha J. P. (2012) *Medical Information Extraction: Information Extraction from Portuguese Hospital Discharge Letters*, Saarbrücken, Germany: Lambert Academic Publishing.

³⁶ <http://www.popline.org>

³⁷ <https://clinicalinformatics.stanford.edu/>

- Gonzalez, R. B. (2008) 'Index Compression for Information Retrieval Systems', Ph.D. Thesis, University of A Coruña.
- Gurulingappa, H., Toldo, L., Rajput, A. M., Kors, J. A., Taweel, A. & Tayrouz, Y. (2013) 'Automatic detection of adverse events to predict drug label changes using text and data mining techniques', *Pharmacoepidemiology Dr S*, 22:1189–1194.
- Hakenberg, J., Voronov, D., Nguyễn, V. H., Liang, S., Anwar, S., Lumpkin, B., Leaman, R., Tari L. & Baral, C. (2012) 'A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions', *J Biomed Inform*, 45:842–850.
- Hall, A. & Walton, G. (2004) 'Information overload within the health care system: a literature review', *Health Inform Libr J*, 21:102–108.
- Herrera-Viedma, E. (2001) 'Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach', *J Am Soc Inform Sci Tech*, 52(6):460–475.
- Imambi, S. S. & Sudha, T. (2010) 'Building classification system to predict risk factors of diabetic retinopathy using text mining', *Int J Comput Sci Eng*, 2(7):2309–2312.
- Imambi, S. S. & Sudha, T. (2011) 'Classification of Medline documents using global relevant weighting schema', *Int J Comput Appl*, 16(3):45–48.
- Jensen, P. B., Jensen, L. J. & Brunak, S. (2012) 'Mining electronic health records: towards better research applications and clinical care', *Nat Rev Gen*, 13:395–405.
- Kankar, P., Adak, S., Sarkar, A. & Sharma, G. (2002) 'MedMeSH Summarizer: Text mining for gene clusters', *Proceedings of the Second SIAM International Conference on Data Mining*.
- Kim, S., Kwon, D., Shin, S.-Y. & Wilbur, W. J. (2012) 'PIE the search: searching PubMed literature for protein interaction information', *Bioinformatics*, 28:597–598.
- Koopman, B., Bruza, P. D., Sitbon, L. & Lawley, M. (2011) 'Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval', *Proceedings of the 1st Australian Workshop on Artificial Intelligence in Health (AIH 2011)*, pp. 1–11.
- Krauthammer, M. & Nenadic, G. (2004) 'Term identification in the biomedical literature', *J Biomed Inform*, 37(6):512–526.
- Kupershmidt, I., Qiaojuan, J. S., Grewal, A., Sundaresh, S., Halperin, I., Flynn, J., Shekar, M., Wang, H., Park, J., Cui, W., Wall, G. D., Wisotzkey, R., Alag, S., Akhtari, S. & Ronaghi, M. (2010) 'Ontology-based meta-analysis of global collections of high-throughput public data', *PLoS ONE*, 5.
- Latha, K., Kalimuthu, S. & Rajaram, R. (2007) 'Information extraction from biomedical literature using text mining framework', *IJISE*, GA, USA, 1(1):1–5.
- Lehner, W., Soderland, S., Aronow, D., Feng, F. & Shmueli, A. (1995) 'Inductive text classification for medical applications', *J Exp Theor Artif In*, 7(1):49–80.
- Li, Y. H. & Jain, A. K. (1998) 'Classification of text documents', *Comput J*, 41(8).
- Liu, Z. & Chu, W. W. (2007) 'Knowledge-based query expansion to support scenario-specific retrieval of medical free text', *Inform Ret*, 10:173–202.
- Liu, H., Lussier, Y. A. & Friedman, C. (2001) 'Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method', *J Biomed Inform*, 34:249–261.
- Lovis, C., Baud, R. H. & Planche, P. (2000) 'Power of expression in the electronic patient record: Structured data or narrative text?' *Int J Med Inform*, 58–59:101–110.
- Manning, C., Raghavan, P. & Schütze, H. (2008) *Introduction to Information Retrieval*, Cambridge: Cambridge University Press.

- Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. (2008) 'Extracting information from textual documents in the electronic health record: a review of recent research', *Yearb Med Inform*, pp. 128–144.
- Mitchell, T. M. (1997) '*Machine Learning*', New York: McGraw-Hill.
- Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. & Tsujii, J. (2006) 'Semantic retrieval for the accurate identification of relational concepts in massive text bases', In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL'06*. pp. 1017–1024.
- Moumtzoglou, A. & Kastania, A. (2011) 'E-Health systems quality and reliability: models and standards', *Medical Information Science Reference*. New York: Hershey.
- Nahm, U. Y. & Mooney, R. J. (2002) 'Text Mining with Information Extraction', *AAAI Tech Rep SS-02-06*, pp. 60–67.
- Nunes, T., Campos, D., Matos, S. & Oliveira, J.L. (2013) 'BeCAS: b Quinlan, Biomedical concept recognition services and visualization', *Bioinformatics*, vol. 29, no. 15, p. 1915–1916, June 2013.
- National Center for Biotechnology Information 2010 PubMed stop words.
- Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J. & Ananiadou, S. (2008) 'Kleio: a knowledge-enriched information retrieval system for biology', In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 787–788.
- Nunes, T., Campos, D., Matos, S. & Oliveira, J. L. (2013) 'BeCAS: biomedical concept recognition services and visualization', *Bioinformatics*, 29:1915–1916.
- Quinlan, J. (1993) '*C4.5: Programs for machine learning*', Morgan Kaufmann: San Matteo, CA.
- Quinlan, J. R. (1998) 'Mini boosting decision trees', *J Artif Intell Res*, 1–15.
- Ramampiaro (2010) 'Retrieving biomedical information with BioTracer: Challenges and possibilities', *NIK-2009*.
- Rebholz-Schuhmann, D., Oellrich, A. & Hoehndorf, R. (2012) 'Text-mining solutions for biomedical research: enabling integrative biology', *Nat Rev Gen*, 13:829–839.
- Simpson, M. S. & Demner-Fushman, D. (2012) 'Biomedical text mining: a survey of recent progress', In C. C. Aggarwal and C. X. Zhai (eds.), *Mining Text Data*, Heidelberg: Springer Verlag, pp. 465–517.
- Singhal, A. (2001) 'Modern information retrieval: a brief overview', *IEEE Data Eng Bull*, 24(4):35–43.
- Srinivasan, P., Bisharah, L. & Sehgal, A. (2004) 'Mining MEDLINE: Postulating a beneficial role for curcumin longa in retinal diseases', Boston, MA: *Workshop: Biolink, Linking Biological Literature, Ontologies and Databases*, pp. 33–40.
- Srinivasan, P. & Libbus, B. (2004) 'Mining MEDLINE for implicit links between dietary substances and diseases', *Bioinformatics*, 20:290–296.
- Suominen, H. (2009) 'Machine learning and clinical text: Supporting health information flow', *TUCS Dissertations*, (125).
- Thoroddsen, A., Saranto, K., Ehrenberg, A. & Sermeus, W. (2009) 'Models, standards and structures of nursing documentation in European countries', *Stud Health Tech Inform*, 146:327–331.
- Van Rijsbergen, C. J. (1979) *Information Retrieval*, 2nd edition, Newton, MA: Butterworth Heinemann.
- Vapnik, V. (1995) '*The Nature of Statistical Learning Theory*', 2nd edition, Heidelberg, Germany: Springer-Verlag. pp. 138–141.

- Vazquez, M., Krallinger, M., Leitner, F. & Valencia, A. (2011) 'Text mining for drugs and chemical compounds: Methods, tools and applications', *Molecular Informatics*, 30:506–519. Available at: <http://doi.wiley.com/10.1002/minf.201100005>.
- Walsh, S. H. (2004) 'The clinician's perspective on electronic health records and how they can affect patient care', *Br Med J*, 328:1184–1187.
- Wang, J., Chen, Q. & Chen, Y. (2004) 'RBF kernel based Support Vector Machine with universal approximation and its application', In *Lecture Notes in Computer Science 3174*, F. Yin, J. Wang, & C. Guo (eds.), Springer Verlag: Heidelberg.
- Yuan, L. (2010) 'An improved Naive Bayes text classification algorithm in Chinese information processing', *Proceedings of the Third International Symposium on Computer Science and Computational Technology (ISCST '10)*.
- Zanasi, A. (2009) 'Virtual weapons for real wars: Text mining for national security', *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08, Advances in Soft Computing*, 53:53–60.
- Zhang, Y, Chen, J. & Xiong (2007) 'Improved Naive Bayes text classification algorithm', *J Guangxi Normal University (Natural Science Edition)*, 2.
- Zhao, L.-L., Zhang, T., Zhuang, L.-W., Yan, B.-Z., Wang, R.-F. & Liu, B.-R. (2014) 'Uncovering the pathogenesis and identifying novel targets of pancreatic cancer using bioinformatics approach', *Mol Biol Rep*, 1–8.
- Zhou, G., Zhang, J., Su, J., Shen, D. & Tan, C. L. (2004) 'Recognizing names in biomedical texts: a machine learning approach', *Bioinformatics*, 20:1178–1190.
- Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W. & Shen, B. (2013) 'Biomedical text mining and its applications in cancer research', *J Biomed Inform*, 46:200–211.

Hua Xu and Joshua C. Denny

2 Unlocking information in electronic health records using natural language processing: a case study in medication information extraction

Abstract: Clinical natural language processing (NLP), which can unlock detailed patient information from clinical narratives stored in electronic health records, has been frequently used to support clinical research and operations. This chapter introduces the state-of-the-art work in clinical NLP. Using medication information extraction as a use case, we describe different methods to build clinical NLP systems, including rule-based, machine learning-based, and hybrid approaches. Applications of medication information extraction systems, such as *pharmacovigilance* (post-market surveillance of drugs) research, are also discussed in this chapter.

2.1 Introduction to clinical natural language processing

Electronic health record (EHR) systems have been increasingly adopted in the United States and worldwide (Jha et al. 2009; Shea and Hripcsak 2010). This growth is fueled, in part, by recent federal legislation that provides significant financial incentives to institutions demonstrating aggressive application and “meaningful use” of comprehensive EHRs (<http://www.hhs.gov/news/press/2010pres/07/20100713a.html>). The ever-growing availability of EHR data has become an enabling resource for clinical and translational research (Kohane 2011). However, the majority of EHR data is narrative text, given that clinical documentation is the primary form of communication in clinical practice. Unstructured clinical texts contain rich patient information, though such texts are not immediately accessible to computerized applications that rely on structured inputs, such as decision support systems and healthcare analytic tools. As a result, there has been a great interest in developing clinical natural language processing (NLP) methods to unlock information embedded in clinical narratives (Meystre et al. 2008; Nadkarni et al. 2011).

Various clinical NLP systems have been developed in past decades to extract information from clinical narratives to facilitate patient care and clinical research. The Linguistic String Project (LSP) (Sager et al. 1987, 1994) led by Naomi Sager at New York University was one of the earliest attempts to formulate comprehensive semantic and syntactic rules to parse clinical text. Later, Friedman and her colleagues (1994) developed a clinical NLP system called MedLEE (Medical Language Extraction and Encoding System), which was originally designed for decision-support applications in the domain of radiology reports of the chest. MedLEE has been shown to be as accurate as physicians at extracting clinical concepts from chest radiology reports (Hripcsak et al. 2002). It is routinely used to process and encode clinical text at New York Presbyterian Hospital. MPLUS and its ancestor SymText (Haug et al. 1995) are NLP systems developed at the University of Utah, which have been used for various applications such as encoding chief complaints into ICD-9 codes and extracting pneumonia-related findings from chest radiograph reports (Fizman et al. 2000). KnowledgeMap Concept Indexer (Denny et al. 2003), an NLP system developed at Vanderbilt University Medical Center (VUMC) around 2000, has been used at Vanderbilt to extract clinical concepts from clinical documents (Denny et al. 2005). Other research groups have also developed various NLP systems for processing clinical text in different sub-domains of medicine (Hahn et al. 2002; Zeng et al. 2006; Harkema et al. 2009; Yetisgen-Yildiz et al. 2013).

Two widely used clinical NLP systems that are freely available to the public are MetaMap and cTAKES (clinical Text Analysis and Knowledge Extraction System). MetaMap (Aronson 2001; Aronson and Lang 2010) is a general biomedical NLP system developed by Aronson et al. at National Library Medicine. It was originally developed to map biomedical literature (e.g., MEDLINE abstracts) to concepts in the Unified Medical Language System (UMLS) Metathesaurus. Many researchers have used MetaMap to extract information from clinical text (Schadow and McDonald 2003; Chung and Murphy 2005; Meystre and Haug 2006; Friedlin and Overhage 2011). For example, Meystre and Haug (2006) applied MetaMap to extract medical problems from clinical text and reported a recall of 0.74 and a precision of 0.76. cTAKES (Savova et al. 2010b) is another freely available comprehensive clinical NLP system, which is developed on the Unstructured Information Management Architecture (UIMA, <http://uima.apache.org/>) framework and the OpenNLP toolkit (<http://opennlp.apache.org/>). cTAKES is a pipeline-based system that consists of different modules such as sentence boundary detector, part-of-speech tagger, shallow parser, and named entity recognizer. Many studies have reported the use of cTAKES for different clinical information extraction tasks such as determining patient smoking status (Savova et al. 2008) and identifying disease cohorts (Savova et al. 2010a).

General-purpose clinical NLP systems such as MedLEE, MetaMap, and cTAKES are often comprehensive, requiring different methodologies for various components. To better describe methods in clinical NLP research, we decided to focus on a more narrowly defined topic. In this chapter, we will use medication information extraction as a use case to explain how state-of-the-art clinical NLP systems work.

2.2 Medication information in EHRs

The use of computer applications for recording and processing drug information are becoming increasingly available in most EHRs. For the inpatient setting, computerized provider order entry (CPOE) systems and electronic medication administration record (eMAR) systems have been widely adopted. Many EHR systems have also incorporated e-prescribing systems in the outpatient setting, which create structured records during generation of new prescriptions and refills. Adoption is increasing, as Meaningful Use Stage 1 requires that 40% of permissible prescriptions are generated and sent to pharmacies electronically by e-prescribing tools. Nevertheless, e-prescribing tools are still not yet widely adopted by physicians. Furthermore, it is often the case that historical medication information is not even generated through the use of such tools. Currently, outpatient medication information is frequently recorded via narrative text entries within clinical documentation or patient problem lists. Not surprisingly, many times this information is transmitted in the course of communications with the patient through telephone calls or patient portals for which there is corresponding notation in the patient's file. For all these reasons given above, an accurate construction of a patient's medication exposure history often requires extraction of information embedded in clinical narratives.

Figure 2.1 shows an example of an outpatient clinic visit note, with medication information highlighted using the underline. As shown in the example, some medication mentions are recorded in a semi-structured list (e.g., in the MEDICATIONS section); while other medications are recorded in narrative sentences (e.g., in the ASSESSMENT AND PLAN section). In the MEDICATION section, a medication entry often contains the medication name (generic or brand) and its signature information, such as dose, form, route, frequency and, sometimes, the reason(s) for giving the patient the medication. Though context-specific information such as reasons or duration of a medication are also important, it is often much more challenging to extract from the patient's record. To complicate things further, medications are often mentioned in the

Chief complaint: SOB and chest pain

History of present illness:

Mrs. X is a 53 year old female with h/o DM2, htn, HLD, prior CAD s/p drug eluting stent 2 months ago who presents with acute onset chest pain earlier today radiating to the left arm and back. She describes it as a strong pressure with SOB but no diaphoresis. It began around 4 am and awoke her from sleep. She then took 2 SL NTG, which ameliorated her pain. No nausea or vomiting. She also describes that she has been having similar chest pains and dyspnea with exertion over the last few months. This has been getting worse in severity. She says her DM has recently not been well controlled. Her last hb A1c was 12 last month. She takes daily ASA but is not currently taking Plavix.

Past medical history:

- Hyperlipidemia
- CAD
- Diabetes mellitus type 2

...

Medications:

- Nexium 40 mg Cap 1 capsule by mouth daily for GERD
- Amoxicillin 1000 mg tablets 1 tablet by mouth three times daily for seven day(s) for acute sinusitis
- Ambien 5 mg qhs prn sleep
- Simvastatin 20 mg Tab (Zocor) 2 tablets by mouth qhs
- Aspirin 325 mg daily
- Metformin 1000 mg bid

Allergies:

- PCN – rash
- ACEI – angioedema

Family medical history:

- No family history of diabetes. mother has breast cancer and was on tamoxifen. Father has had lung CA and died of an MI at 64.

...

Assessment and plan:

1. Will admit with to cardiology service for cardiac catheterization ...
2. DMII: hb A1c is elevated. will hold metformin and change to SSI.
3. HTN: improve BP control. will add beta blocker and ARB (since allergic to ACEI). continue ASA.

Fig. 2.1: An example of outpatient clinic visit note, where medication information is highlighted with underline.

patient's file for other purposes than for treating the medical condition at hand. For example, as shown in Fig. 2.1, medications are mentioned in the patient's file for a variety of reasons: (1) to indicate possible allergies or adverse reactions; (2) to formulate a family medical history: the name of the medication taken by the patient's mother is useful for defining the exact kind of breast cancer the mother had; and (3) medications that the patient is *not* taking may

be mentioned in the “history of present illness” section of patient’s chart to indicate possible lapse in medical care that must be addressed. For example, the blood thinner “Plavix” (the brand name for clopidogrel) for treating cardiac problems appears in this section of the patient’s record.

2.3 Medication information extraction systems and methods

2.3.1 Relevant work

Early studies on medication information extraction in EHRs have been focused on identifying drug names and selected signature information such as dosage. For example, Chhieng et al. (2007) used a string-matching method to identify drug names in clinical records and reported a precision rate of 83%. Levin and colleagues (2007) extracted drug names from anesthesia records and reported high performance with a sensitivity of 92.2% and a specificity of 95.7%. Evans et al. (1996) developed the CLARIT system and showed that it could extract drug name and dosage phrases in patient discharge summaries with an accuracy of 80%.

More recent studies extended the scope to additional drug signature information such as *route* and *frequency*. Gold et al. (2008) developed a regular expression based approach to extracting drug names and signature information including dose, route, and frequency. They evaluated the system using a data set of 26 discharge summaries and showed that drug names were identified with a precision of 94.1% and a sensitivity of 82.5%, but other signature information such as dose and frequency had much lower precisions. In a study by Jagannathan et al. (2009), several commercial systems, such as LifeCode™, FreePharma™, and Coderyte, were assessed for their ability to extract medication information (including drug names, strength, route, and frequency) from clinical notes. Their evaluation showed a high F-measure of 93.2% on capturing drug names, but lower F-measures of 85.3%, 80.3%, and 48.3% on retrieving strength, route, and frequency, respectively (Jagannathan et al. 2009). At VUMC, a medication information extraction system called MedEx (Xu et al. 2010) was developed. It achieved F-scores over 90% on extracting drug names, strength, route and frequency information in discharge summaries and clinic visit notes from VUMC’s EHR.

In 2009, i2b2 (Center of Informatics for Integrating Biology and Beside) organized a clinical NLP challenge to extract medication-related information in discharge summaries from Partners Healthcare (Uzuner et al. 2010). The goal of

the challenge was to identify and determine boundaries of six types of drug information, which consisted of 1) drug name, 2) dosage, 3) route, 4) frequency, 5) duration; and 6) reason for drug administration. In addition, it required determination of whether medication information was found in a list or a narrative sentence. Figure 2.2 (Uzuner et al. 2010) shows the examples of inputs and outputs of the 2009 i2b2 challenge. Twenty teams, including international entries, participated in the medication challenge using various approaches including rule-based, machine learning and hybrid methods (see Section 2.3.2) (Deleger et al. 2010; Doan et al. 2010; Hamon and Grabar 2010; Li et al. 2010; Meystre et al. 2010; Mork et al. 2010; Patrick and Li 2010; Spasic et al. 2010; Tikk and Solt 2010; Yang 2010). While names of medications were well identified by all of the top 10 systems, the performance for durations and reasons were still low, with the best F-measure of 0.525 and 0.459, respectively (Uzuner et al. 2010).

The i2b2 medication challenge created an important asset to enhance research in this area by generating an annotated dataset for medication information extraction in EHRs. Using the i2b2 data set, researchers investigated different aspects of machine learning-based approaches to medication-entity recognition, including different machine learning algorithms (Doan 2010) and the ensemble method, which combines predictions from multiple classifiers (Doan et al. 2012). Furthermore, Li et al. (2013) extended such medication information extraction methods to other clinically relevant text such as clinical trial documents.

Line no. text	
63	well. Although left transmetatarsal amputation being considered,
64	it was felt that she had a good chance of healing the wound
65	appropriately. She had a single temperature spike, although all
66	cultures remained negative. <i>She had continuation of her Heparin</i>
67	<i>while she was started on a course of Coumadin to reserve patency of</i>
68	<i>her graft. ...</i>
Gold standard	
m="heparin" 66:8 66:8 do="nm" mo="nm" f="nm" du="nm" r="nm" ln="narrative"	
m="coumadin" 67:8 67:8 do="nm" mo="nm" f="nm" du="nm" r="her graft." 68:0	
68:1 ln="narrative"	

Fig. 2.2: Examples of the input and output in the 2009 i2b2 medication information extraction challenge. The challenge requires to identify six types of drug information including drug name (m), dosage (do), route (mo), frequency (f), duration (du), and reason (r), as well as their exact offset (by line number and token position) in the clinical documents. The figure was taken from (Uzuner et al. 2010).

2.3.2 Summary of approaches

Although many systems have been developed to extract medication information from clinical text, their methodological approaches can be mainly divided into three categories: rule-based (Gold et al. 2008; Deleger et al. 2010; Mork et al. 2010; Spasic et al. 2010; Xu et al. 2010), machine learning-based (Li et al. 2010; Patrick and Li 2010; Li et al. 2013), and hybrid methods (Meystre et al. 2010; Tikk and Solt 2010).

2.3.2.1 Rule-based methods

A rule-based medication information extraction system often works as following: (1) identify medication-related entities by using rules and dictionaries (e.g., lists of drug names) based on domain-specific resources; (2) filter medication entries based on context-specific information; and (3) link signature modifiers to corresponding drug names using specific rules.

Medication name identification is the crucial step in medication information extraction. Rule-based systems often leverage existing medical knowledge to build a comprehensive list of drug names. In Mork et al. (2010), the drug dictionary was built on various medical resources such as UMLS (<http://www.nlm.nih.gov/research/umls>), RxNorm (<http://www.nlm.nih.gov/research/umls/rxnorm>), and DailyMed (<http://dailymed.nlm.nih.gov>). While lexicons for some signature fields such as route could be built in a similar way by creating an exhaustive list; other signature fields such as dose and frequency have to be recognized by defining regular expression patterns. Figure 2.3 shows some examples of regular expression rules used in Yang's (2010) system for recognizing frequency expressions. In the system developed by Spasic and colleagues (2010), all rules were implemented as expression in Mixup (My Information eXtraction and Understanding Package), a simple pattern-matching language.

After a possible drug entity is recognized, context-based rules have to be applied to verify its inclusion. A drug entry could be excluded in the final output due to several reasons, such as drug-allergy information, negation, and non-patient experience (e.g., about a family member). Specific rules could be developed based on context information, e.g., to remove drugs in the Allergy section. In addition, drug names could be ambiguous as well. For example, in the sentence “The patient was found to be iron deficient and she continued on iron supplements,” the first occurrence of “iron” should not be labeled as a medication, as it is associated with “deficient,” a term that is used here to indicate a medical problem in this particular patient. However, the second occurrence of “iron” should be marked as a *medication*, because it is collocated with the word “supplements.” What we learn from this is that it is very important to construct rules to resolve such ambiguities so that we can improve the performance of the medication information extraction system.

Token-based rule	Example
[after before at following with w/] + <Meal>	after breakfast, before meals, at supper, following lunch
[in on at during]+<Daytime>	in the a.m., at bedtime, on p.m., during the evening
[each every on]+<Weekday>	each Monday, every Sunday, on tues,
[every]+<Num>+<TimeUnit>	every 3 hour, every 3–5 min
<Num>+[x x/]+<TimeUnit>	2×/wk, 2–3×/day, 2×wk
[q q.] +<TimeUnit>	qhr, q day, q.wk, q. week
[q q.] +<Num>+<TimeUnit>	q2h, q 4 h, q. 2 weeks, q.6 h
[q q.] +<Meal>	qlunch, q breakfast, q.meal, q. dinner
[q q.] +<Daytime>	qam, q p.m., q. afternoon, q.evening
[q q.] +<Weekday>	qwed, q monday, q. friday, q.saturday
[once twice] + <OneTimeUnit>	once a day, twice per day
<Num>+[times x] +<OneTimeUnit>	2 times a day, 3×daily

Fig. 2.3: Examples of regular expressions for recognizing frequency expressions, as specified in Yang (2010).

The last step is to link drug names with their corresponding signature modifiers. A simple but effective approach is to use regular expression to recognize drug names together with their associated signature fields, as implemented in the MERKI system developed by (Gold et al. 2008). However, sometimes clinical sentences could be very complex, containing multiple medications with repetitive signature text, e.g., “*Midrin 2 po initial then 1 po q6hrs prn.*” To better handle such complex cases, Xu et al. (2010) developed a more robust approach that uses a chart parser and a semantic grammar to parse medications in a sentence based on a formal representation model.

2.3.2.2 Machine learning-based methods

From the perspective of supervised machine learning, the medication information extraction task in the 2009 i2b2 challenge can be divided into two steps: (1) identifying medication-related entities (e.g., drug names and other signature fields); and (2) determining the linkage between the detected medication names and the other signature modifiers. Both tasks can be converted into classification problems and resolved, using supervised machine-learning approaches.

Identification of medication-related entities is a typical named entity recognition (NER) problem, which is to determine boundary and semantic classes (e.g., medication, dosage, or frequency) of words/phrases in free text. To apply machine learning algorithms to an NER task, annotated text are typically converted into a “BIO” format. Specifically, it assigns each token into one of the three classes: **B** – beginning of an entity, **I** – inside an entity, and **O** – outside of an entity. Thus, an NER problem now is converted to a classification problem – to determine a correct label {B, I, O} for each token. Figure 2.4 shows a clinical sentence and its corresponding BIO labels. As multiple types of drug-related entities (e.g., drug name, dose, and frequency) need to be identified, we can extend the BIO labels by adding a suffix to indicate its entity type. For example, “B-m” indicates the beginning of a medication entity, “B-d” indicates the beginning of a dose entity, and “B-f” is used to indicate the beginning of a frequency entity. Different machine learning algorithms have been used for NER tasks. For example, Conditional Random Fields (CRFs) (Lafferty et al. 2001), a representative model for sequence labeling, is one of the most widely used algorithms. In the studies by Patrick and Li (2010) and Li et al. (2010), CRF was used to recognize medication-related entities. Doan and Xu (2010) developed a medication entity recognition approach using Support Vector Machines (SVMs) and reported reasonable performance as well.

As mentioned above, machine learning-based approaches can also be implemented to determine the linkage between the detected medication names and signature modifiers. An intuitive approach to linkage detection would be to build a binary classifier to determine if a candidate pair of medication name and signature modifier is linked or not. Candidate linkage pairs can be generated by taking all possible medication name and signature modifier pairs in one sentence. For example, Patrick and Li (2010) developed an SVM-based classifier to determine the linkage between drug names and signature modifiers in the 2009 i2b2 challenge. Li et al. (2013) proposed a multi-layered sequence labeling for medication-signature linkage detection. However, their evaluation showed that the multi-layered sequence labeling approach did not perform as well as the SVM-based binary classifier.

Token:	In	addition	,	start	Percocet	1-2	tablets	twice	a	day
Label:	O	O	O	O	B-m	B-d	I-d	B-f	I-f	I-f

Fig. 2.4: An example of the BIO representation of an annotated clinical sentence. Upper case letters {B, I, O} stand for beginning of an entity (B), inside an entity (I), and outside of an entity (O) respectively. Lower case letters {m, d, f} stand for entity types, medication name (m), dose (d), and frequency (f).

2.3.2.3 Hybrid methods

A hybrid system takes advantages of both rule-based methods and machine learning-based methods. It often consists of modules that are based on machine learning algorithms and modules that use regular expressions, rules, and dictionaries. Different approaches have been developed to combine machine learning and rule-based methods for medication information extraction. In Patrick and Li's system (Patrick and Li 2010), context-specific rules were applied to outputs of machine learning-based modules in a post-processing fashion. Others used outputs of rule-based modules to improve machine learning classifiers. For example, Doan and Xu (2010) used outputs of a rule-based system as features to feed into a machine learning classifier and demonstrated improved performance. Tikk and Solt (2010) used a rule-based system to create additional training datasets for a machine learning system and also showed better performance.

2.4 Uses of medication information extraction tools in clinical research

Practice-based structured medication data (e.g., claims) have long been used for a large variety of drug outcome studies, including pharmacoepidemiology, pharmaco-economic, and service-related healthcare investigations (Strom 2005). EHR data, which can include more comprehensive lists of patients' drug exposure (including both over-the-counter and prescription medications) and clinical outcomes, have emerged as a new enabling resource to facilitate broad types of drug-related clinical studies, including pharmacovigilance (Wang et al. 2009) and pharmacogenomics (Wilke et al. 2011). All such studies rely on medication information extraction tools to automatically and accurately extract patient medication exposure information from EHRs.

Pharmacovigilance: Post-market surveillance (also called pharmacovigilance) is an important step to establish complete safety profiles of drugs by detecting additional adverse drug reactions (ADRs) that are not captured during clinical trial phases. Current pharmacovigilance databases such as US Food and Drug Administration's Adverse Event Reporting System (FAERS) have limitations. As a result, EHRs are emerging as a promising new data source for pharmacovigilance (Wood and Martinez 2004; Wysowski and Swartz 2005). Wang et al. (2009) conducted a feasibility study that used the MedLEE system (Friedman et al. 1994) to extract medication and adverse drug events from hospital discharge summaries and then calculated co-occurrence statistics between these events. Their evaluation showed it was feasible to detect known drug ADRs, as well as novel ADRs from EHRs. The same group then applied

similar informatics methods to detect two serious ADRs: rhabdomyolysis and agranulocytosis from EHRs, and showed promising results (Haerian et al. 2012). More recently, La Pendu et al. (2013), a group of researchers at Stanford University, also demonstrated the use of EHRs and NLP methods to conduct pharmacovigilance studies, including detecting ADRs associated with drug-drug interactions.

Pharmacogenomics: Recently, huge efforts have been initiated to link new and existing EHR databases with archived biological material, to accelerate research in personalized medicine, such as pharmacogenomics that aims to identify common and rare genetic variants that contribute to variability in drug response specifically within the context of relevant clinical covariates (McCarty and Wilke 2010). One such effort has been the NIH-funded eMERGE network (*electronic MEDical Records and GENomics*) (Manolio 2009), a consortium of institutions with DNA biobanks coupled with large comprehensive EHRs. For example, the research team at Vanderbilt has used a DNA biobank linked to de-identified EHR data to successfully replicate pharmacogenomics associations between cardiovascular risk and *CYP2C19*2* and *ABCB1* in patients receiving clopidogrel (Delaney et al. 2012). They also looked at associations between variants in *CYP2C9*, *VKORC1* and *CYP4F2* and a steady state Warfarin (a blood thinner) dose in individuals of European and African ancestry (Ramirez et al. 2012). In both studies, MedEx (Xu et al. 2010) was used to help identify drug exposure information of patients in EHRs. In addition, the Vanderbilt team also extended MedEx to automatically extract weekly dose of Warfarin (Xu et al. 2011) and daily dose of statins (Wei et al. 2014) from EHRs to facilitate pharmacogenomic studies of both drugs.

2.5 Challenges and future work

Clinical NLP has become an enabling technology to unlock unstructured data in EHRs to support the secondary use of EHR data for clinical and translational research. This chapter briefly introduces relevant work, methods, and applications of clinical NLP technologies, using the task of medication information extraction as an example. Although the content is specific for medication information in EHRs, the NLP methods described here are generalizable to other types of clinical information found in EHRs.

Despite the promising results achieved by current medication information extraction systems, it is still challenging to accurately extract contextual information for medications, such as duration and reason of medication administration (Uzuner et al. 2010). These types of context are often loosely attached

with medication mentions, e.g., the reason may be located in a different sentence than the drug name and using a variety of different words to indicate the linkage, if specified at all. In such situations, sentence syntactic structures could be helpful though parsing clinical text is still an under-explored area of clinical NLP. In addition, existing knowledge bases about drug and indication pairs could potentially be helpful. Existing resources, such as SIDER (Kuhn et al. 2010) and MEDI (Wei et al. 2013), may allow a hybrid approach to improved drug-indication linkage. All in all, integrating domain specific knowledge with machine learning-based information extraction systems remains a challenging task (Friedman et al. 2013).

Another exciting direction of future work, described by Liu and her colleagues (2011) is to build longitudinal drug profiles of patients (e.g., to link longitudinal drug mentions of the same patient to form a timeline of drug exposure events such as the “start” or “discontinuation” of a drug), which is important for any drug-related clinical study.

References

- Aronson, A. R. (2001) ‘Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program’, *Proc AMIA Symp*, 17–21.
- Aronson, A. R. & Lang, F. M. (2010) ‘An overview of MetaMap: historical perspective and recent advances’, *J Am Med Inform Assoc*, 17:229–236.
- Chhieng, D., Day, T., Gordon, G. & Hicks, J. (2007) ‘Use of natural language programming to extract medication from unstructured electronic medical records’, *AMIA Annu Symp Proc*, 908.
- Chung, J. & Murphy, S. (2005) ‘Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports’, *AMIA Annu Symp Proc*, 131–135.
- Delaney, J. T., Ramirez, A. H., Bowton, E., Pulley, J. M., Basford, M. A., Schildcrout, J. S., Shi, Y., Zink, R., Oetjens, M., Xu, H., Cleator, J. H., Jahangir, E., Ritchie, M. D., Masys, D. R., Roden, D. M., Crawford, D. C. & Denny, J. C. (2012) ‘Predicting clopidogrel response using DNA samples linked to an electronic health record’, *Clin Pharmacol Ther*, 91:257–263.
- Deleger, L., Grouin, C. & Zweigenbaum, P. (2010) ‘Extracting medical information from narrative patient records: the case of medication-related information’, *J Am Med Inform Assoc*, 17:555–558.
- Denny, J. C., Smithers, J. D., Miller, R. A. & Spickard, A., 3rd (2003) ‘“Understanding” medical school curriculum content using KnowledgeMap’, *J Am Med Inform Assoc*, 10:351–362.
- Denny, J. C., Spickard, A., 3rd, Miller, R. A., Schildcrout, J., Darbar, D., Rosenbloom, S. T. & Peterson, J. F. (2005) ‘Identifying UMLS concepts from ECG Impressions using KnowledgeMap’, *AMIA Annu Symp Proc*, 196–200.
- Doan, S., Bastarache, L., Klimkowski, S., Denny, J. C. & Xu, H. (2010) ‘Integrating existing natural language processing tools for medication extraction from discharge summaries’, *J Am Med Inform Assoc*, 17:528–531.
- Doan, S., Collier, N., Xu, H., Pham, H. D. & Tu, M. P. (2012) ‘Recognition of medication information from discharge summaries using ensembles of classifiers’, *BMC Med Inform Decis Mak*, 12:36.

- Doan, S. X. H. (2010) 'Recognizing Medication related Entities in Hospital Discharge Summaries using Support Vector Machine'. *Coling 2010, The 23rd International Conference on Computational Linguistics* Beijing.
- Evans, D. A., Brownlow, N. D., Hersh, W. R. & Campbell, E. M. (1996) 'Automating concept identification in the electronic medical record: an experiment in extracting dosage information', *Proc AMIA Annu Fall Symp*, 388–392.
- Fiszman, M., Chapman, W. W., Aronsky, D., Evans, R. S. & Haug, P. J. (2000) 'Automatic detection of acute bacterial pneumonia from chest X-ray reports', *J Am Med Inform Assoc*, 7:593–604.
- Friedlin, J. & Overhage, M. (2011) 'An evaluation of the UMLS in representing corpus derived clinical concepts', *AMIA Annu Symp Proc*, 2011:435–444.
- Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J. & Johnson, S. B. (1994) 'A general natural-language text processor for clinical radiology', *J Am Med Inform Assoc*, 1:161–174.
- Friedman, C., Rindflesch, T. C. & Corn, M. (2013) 'Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine', *J Biomed Inform*, 46:765–773.
- Gold, S., Elhadad, N., Zhu, X., Cimino, J. J. & Hripcsak, G. (2008) 'Extracting structured medication event information from discharge summaries', *AMIA Annu Symp Proc*, 237–241.
- Haerian, K., Varn, D., Vaidya, S., Ena, L., Chase, H. S. & Friedman, C. (2012) 'Detection of pharmacovigilance-related adverse events using electronic health records and automated methods', *Clin Pharmacol Ther*, 92:228–234.
- Hahn, U., Romacker, M. & Schulz, S. (2002) 'MEDSYNDIKATE—a natural language system for the extraction of medical information from findings reports', *Int J Med Inform*, 67:63–74.
- Hamon, T. & Grabar, N. (2010) 'Linguistic approach for identification of medication names and related information in clinical narratives', *J Am Med Inform Assoc*, 17:549–554.
- Harkema, H., Dowling, J. N., Thornblade, T. & Chapman, W. W. (2009) 'ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports', *J Biomed Inform*, 42:839–851.
- Haug, P. J., Koehler, S., Lau, L. M., Wang, P., Rocha, R. & Huff, S. M. (1995) 'Experience with a mixed semantic/syntactic parser', *Proc Annu Symp Comput Appl Med Care*, 284–288.
- Hripcsak, G., Austin, J. H., Alderson, P. O. & Friedman, C. (2002) 'Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports', *Radiology*, 224:157–163.
- Jagannathan, V., Mullett, C. J., Arbogast, J. G., Halbritter, K. A., Yellapragada, D., Regulapati, S. & Bandaru, P. (2009) 'Assessment of commercial NLP engines for medication information extraction from dictated clinical notes', *Int J Med Inform*, 78:284–291.
- Jha, A. K., Desroches, C. M., Campbell, E. G., Donelan, K., Rao, S. R., Ferris, T. G., Shields, A., Rosenbaum, S. & Blumenthal, D. (2009) 'Use of electronic health records in U.S. hospitals', *N Engl J Med*, 360:1628–1638.
- Kohane, I. S. (2011) 'Using electronic health records to drive discovery in disease genomics', *Nat Rev Genet*, 12:417–428.
- Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J. & Bork, P. (2010) 'A side effect resource to capture phenotypic effects of drugs', *Mol Syst Biol*, 6:343.
- Lafferty, J., Mccallum, A. & Pereira, F. (2001) 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data', *Proc. 18th International Conf. on Machine Learning*, 282–289.
- Lependu, P., Iyer, S. V., Bauer-Mehren, A., Harpaz, R., Mortensen, J. M., Podchiyska, T., Ferris, T. A. & Shah, N. H. (2013) 'Pharmacovigilance using clinical notes', *Clin Pharmacol Ther*, 93:547–555.

- Levin, M. A., Krol, M., Doshi, A. M. & Reich, D. L. (2007) 'Extraction and mapping of drug names from free text to a standardized nomenclature', *AMIA Annu Symp Proc*, 438–442.
- Li, Q., Zhai, H., Deleger, L., Lingren, T., Kaiser, M., Stoutenborough, L. & Solti, I. (2013) 'A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction', *J Am Med Inform Assoc*, 20:915–921.
- Li, Z., Liu, F., Antieau, L., Cao, Y. & Yu, H. (2010) 'Lancet: a high precision medication event extraction system for clinical text', *J Am Med Inform Assoc*, 17:563–567.
- Liu, M., Jiang, M., Kawai, V. K., Stein, C. M., Roden, D. M., Denny, J. C. & Xu, H. (2011) 'Modeling drug exposure data in electronic medical records: an application to warfarin', *AMIA Annu Symp Proc*, 2011:815–823.
- Manolio, T. A. (2009) 'Collaborative genome-wide association studies of diverse diseases: programs of the NHGRIs office of population genomics', *Pharmacogenomics*, 10:235–241.
- Mccarty, C. A. & Wilke, R. A. (2010) 'Biobanking and pharmacogenomics', *Pharmacogenomics*, 11:637–641.
- Meystre, S. & Haug, P. J. (2006) 'Natural language processing to extract medical problems from electronic clinical documents: performance evaluation', *J Biomed Inform*, 39:589–599.
- Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. (2008) 'Extracting information from textual documents in the electronic health record: a review of recent research', *Yearb Med Inform*, 128–144.
- Meystre, S. M., Thibault, J., Shen, S., Hurdle, J. F. & South, B. R. (2010) 'Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents', *J Am Med Inform Assoc*, 17:559–562.
- Mork, J. G., Bodenreider, O., Demner-Fushman, D., Dogan, R. I., Lang, F. M., Lu, Z., Neveol, A., Peters, L., Shooshan, S. E. & Aronson, A. R. (2010) 'Extracting Rx information from clinical narrative', *J Am Med Inform Assoc*, 17:536–539.
- Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011) 'Natural language processing: an introduction', *J Am Med Inform Assoc*, 18:544–551.
- Patrick, J. & Li, M. (2010) 'High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge', *J Am Med Inform Assoc*, 17:524–527.
- Ramirez, A. H., Shi, Y., Schildcrout, J., Delaney, J. T., Xu, H., Oetjens, M., Zuvich, R., Basford, M., Bowton, E., Jiang, M., Zink, R., Cowan, J. D., Pulley, J. M., Ritchie, M. D., Peterson, J. F., Masys, D. R., Roden, D. M., Crawford, D. C. & Denny, J. C. (2012) 'Predicting warfarin dosage in European and African Americans using DNA samples linked to an electronic health record'. *Pharmacogenomics*, in press.
- Sager, N., Friedman, C. & Lyman M.S. (1987) *Medical Language Processing: Computer Management of Narrative Data*. Reading, Addison-Wesley: MA.
- Sager, N., Lyman, M., Bucknall, C., Nhan, N. & Tick, L. J. (1994) Natural language processing and the representation of clinical data. *J Am Med Inform Assoc*, 1:142–160.
- Savova, G. K., Fan, J., Ye, Z., Murphy, S. P., Zheng, J., Chute, C. G. & Kullo, I. J. (2010a) 'Discovering peripheral arterial disease cases from radiology notes using natural language processing', *AMIA Annu Symp Proc*, 2010:722–726.
- Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. & Chute, C. G. (2010b) 'Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications', *J Am Med Inform Assoc*, 17:507–513.
- Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D. & Chute, C. G. (2008) 'Mayo clinic NLP system for patient smoking status identification', *J Am Med Inform Assoc*, 15:25–28.

- Schadow, G. & McDonald, C. J. (2003) 'Extracting structured information from free text pathology reports', *AMIA Annu Symp Proc*, 584–588.
- Shea, S. & Hripcsak, G. (2010) 'Accelerating the use of electronic health records in physician practices', *N Engl J Med*, 362:192–195.
- Spasic, I., Sarafraz, F., Keane, J. A. & Nenadic, G. (2010) 'Medication information extraction with linguistic pattern matching and semantic rules', *J Am Med Inform Assoc*, 17:532–535.
- Strom, B. L. (2005) *Pharmacoepidemiology*, J. Wiley: Chichester; Hoboken, NJ.
- Tikk, D. & Solt, I. (2010) 'Improving textual medication extraction using combined conditional random fields and rule-based systems', *J Am Med Inform Assoc*, 17:540–544.
- Uzuner, O., Solti, I. & Cadag, E. (2010) 'Extracting medication information from clinical text', *J Am Med Inform Assoc*, 17:514–518.
- Wang, X., Hripcsak, G., Markatou, M. & Friedman, C. (2009) 'Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study', *J Am Med Inform Assoc*, 16:328–337.
- Wei, W. Q., Cronin, R. M., Xu, H., Lasko, T. A., Bastarache, L. & Denny, J. C. (2013) 'Development and evaluation of an ensemble resource linking medications to their indications', *J Am Med Inform Assoc*, 20:954–961.
- Wei, W. Q., Feng, Q., Jiang, L., Waitara, M. S., Iwuchukwu, O. F., Roden, D. M., Jiang, M., Xu, H., Krauss, R. M., Rotter, J. I., Nickerson, D. A., Davis, R. L., Berg, R. L., Peissig, P. L., Mccarty, C. A., Wilke, R. A. & Denny, J. C. (2014) 'Characterization of statin dose response in electronic medical records', *Clin Pharmacol Ther*, 95(3):331–338.
- Wilke, R. A., Xu, H., Denny, J. C., Roden, D. M., Krauss, R. M., Mccarty, C. A., Davis, R. L., Skaar, T., Lamba, J. & Savova, G. (2011) 'The emerging role of electronic medical records in pharmacogenomics', *Clin Pharmacol Ther*, 89:379–386.
- Wood, L. & Martinez, C. (2004) 'The general practice research database: role in pharmacovigilance', *Drug Saf*, 27:871–881.
- Wysowski, D. K. & Swartz, L. (2005) 'Adverse drug event surveillance and drug withdrawals in the United States, 1969–2002: the importance of reporting suspected reactions', *Arch Intern Med*, 165:1363–1369.
- Xu, H., Jiang, M., Oetjens, M., Bowton, E. A., Ramirez, A. H., Jeff, J. M., Basford, M. A., Pulley, J. M., Cowan, J. D., Wang, X., Ritchie, M. D., Masys, D. R., Roden, D. M., Crawford, D. C. & Denny, J. C. (2011) 'Facilitating pharmacogenetic studies using electronic health records and natural language processing: a case study of warfarin', *J Am Med Inform Assoc*, 18:387–391.
- Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R. & Denny, J. C. (2010) 'MedEx: a medication information extraction system for clinical narratives', *J Am Med Inform Assoc*, 17:19–24.
- Yang, H. (2010) 'Automatic extraction of medication information from medical discharge summaries', *J Am Med Inform Assoc*, 17:545–548.
- Yetisgen-Yildiz, M., Gunn, M. L., Xia, F. & Payne, T. H. (2013) 'A text processing pipeline to extract recommendations from radiology reports', *J Biomed Inform*, 46:354–362.
- Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N. & Lazarus, R. (2006) 'Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system', *BMC Med Inform Decis Mak*, 6:30.

António Teixeira, Liliana Ferreira and Mário Rodrigues

3 Online health information semantic search and exploration: reporting on two prototypes for performing information extraction on both a hospital intranet and the world wide web

Abstract: In this chapter, we apply ontology-based information extraction to unstructured natural language sources to help enable semantic search of health information. We propose a general architecture capable of handling both private and public data. Two of our novel systems that are based on this architecture are presented here. The first system, MedInX, is a Medical Information eXtraction system which processes textual clinical discharge records, performing automatic and accurate mapping of free text reports onto a structured representation. MedInX is designed to be used by health professionals, and by hospital administrators and managers, allowing its users to search the contents of such automatically populated ontologies. The second system, SPHInX, attempts to perform semantic search on health information publicly available on the web in Portuguese. The potential of the proposed approach is clearly shown with usage examples and evaluation results.

3.1 Introduction

More and more healthcare institutions store vast amounts of information about users, procedures, and examinations, as well as the findings, test results, and diagnoses, respectively. Other institutions, such as the Government, increasingly disclose health information on varied topics of concern to the public writ large. Health research is one of the most active areas, resulting in a steady flow of publications reporting on new findings and results.

In recent years, the Internet has become one of the most important tools to obtain medical and health information. Standard web search is by far the most common interface for such information (Abraham & Reddy 2007). Several general search engines such as Google, Yahoo! and Bing currently play an important role in obtaining medical information for both medical professionals and lay persons (Wang et al. 2012). However, these general search engines do not allow the end-user to obtain a clear and organized presentation of the available health

information. Instead, it is more or less of a hit or miss, random return of information on any given search. In fact, medicine-related information search is different from other information searches, since users often use medical terminology, disease knowledge, and treatment options in their search (Wang et al. 2012).

Much of the information that would be of interest to private citizens, researchers, and health professionals is found in unstructured text documents. Efficient access to this information implies the development of search systems capable of handling the technical lexicon of the domain area, entities such as drugs and exams, and the domain structure. Such search systems are said to perform semantic search as they base the search on the *concepts* asked and not so much on the words used in the query (Guha, McCool & Miller 2003). Semantic search maintains several advantages over search based on surface methods, such as those that directly index text words themselves rather than underlying concepts. Three main advantages of concept-based search are: (1) they usually produce smaller sets of results, as they are able to identify and remove semantically duplicated results and/or semantically irrelevant results; (2) they can integrate related information scattered across documents; frequently answers are obtained by compounding information from two or more sources; and (3) they can retrieve relevant results even when the question and answer do not have common words, since these systems can be aware of similar concepts, synonyms, meronyms, antonyms, etc.

Semantic search involves representing the concepts of a domain and the relations between them, organizing, in this way, the information according to its semantics and forming a knowledge representation of the world or some part of it. This representation is called *ontology*, which is formally defined as an explicit specification of a shared conceptualization (Gruber 1993). Ontology describes a hierarchy of concepts related by subsumption relationships, and can include axioms to express other relationships between concepts and to constrain their intended interpretation. The usage of ontology to explicitly define the application domain brings large benefits from the viewpoint of information accessibility, maintainability and interoperability, as it formalizes and allows the application's view of the world to be made public (Guarino 1998). Also, with the emergence of semantic reasoners, software that is able to infer logical consequences from a set of asserted facts or axioms, it is possible to verify the coherence of the stored information and to infer new information from the contents of ontology (Sirin & Parsia 2004).

However, there is still the challenge of bridging the gap between the needed semantically structured information and the original text content. The acquisition of specific and relevant pieces of information from texts, and respective storage in a coherent framework, is called *information extraction* (Cowie & Lehnert 1996). The general problem of information extraction (IE) involves the analysis of natural language texts, such as English or Portuguese texts, to determine the semantic

relations among the existing entities and the events they participate in, namely their relations. Natural language texts can be unstructured, plain texts, and/or semi-structured machine-readable documents, with some kind of markup. The information to be retrieved can be entities, classes of objects and events, and relationships between them. Informally, IE is the task of detecting elements such as “who” did “what” to “whom,” “when” and “where,” in unstructured free text information sources, and using those elements to populate structured information sources (Gaizauskas & Wilks 1998; Màrquez et al. 2008).

IE is different from information retrieval (IR), which is the task usually performed by current search engines such as Google and Bing. Whereas IE aims to extract relevant information from documents, IR aims to retrieve those relevant documents themselves from collections. For example, universities and other public libraries use IR systems to provide access to books, journals and other documents. However, in such cases, after querying search engines the users *still* have to read through those documents brought up in their search to find the information they were looking for. When the goal is to explore data, obtain a summary of facts reported in large amounts of documents or have facts presented in tables, IE becomes a much more relevant technology than IR (McNaught & Black 2006).

A typical IE system has two main subtasks: entity recognition and relation extraction. Entity recognition seeks to locate and classify atomic elements in natural language texts into predefined categories, while relation extraction tries to identify the relations between the entities in order to fill predefined templates. Two important challenges exist in IE. One arises from the variety of ways of expressing the same fact. The other challenge, shared by almost all NLP tasks, is due to the great expressiveness of natural languages, which can have ambiguous structure and meaning.

The chapter is structured as follows: the next two sections provide background information and an overview of related work about information search in general and in the health domain, information extraction in health, and ontology-based information extraction. Section 4 contains the general vision/proposal of an ontology-based information extraction system to feed a search engine. The respective instantiation in two systems with different purposes and some illustrative results are presented in Section 5. Chapter ends with conclusions, provided in Section 6.

3.2 Background

Several approaches to IE have been followed over the years. One common approach is based on pattern matching and exploits basic patterns over a variety of structures: text strings, part-of-speech tags, semantic pairs, and dictionary

entries (Pakhomov 2005). However, this type of approach does not generalize well, which limits its extension to new domains. The need for IE systems that can be easily adapted from one domain to another leads to the development of different approaches based on adaptive IE, starting with the Alembic Workbench (Aberdeen et al. 1995). The idea behind these approaches is to use various kinds of machine learning algorithms to allow IE systems to be easily targeted to new problems. The effort required to redesign a new system is replaced with that of generating batches of training data and applying learning algorithms.

A more recent approach is the ontology-based IE (OBIE), which aims at using ontology to guide the information extraction process (Hahn, Romacker & Schulz 2002). Since Berners-Lee et al. (1994) and Berners-Lee & Fischetti (1999) began to endorse ontologies as the backbone of the semantic web in the 1990s, a whole research field has evolved around the fundamental engineering aspects of ontologies, such as their generation, evaluation and management.

A relevant number of approaches need seed examples to train the IE systems. As such, several tools to annotate the semantic web were developed. Some earlier systems involved having humans annotate texts manually, using user-friendly interfaces (Handschuh, Staab & Studer 2003; Schroeter, Hunter & Kosovic 2003). Others featured algorithms to automate part of the annotation process. Those algorithms were based on manually constructed rules or extraction patterns, to be completed based on the previous annotations (Alfonseca & Manandhar 2002; Ciravegna et al. 2002).

As manual annotation can be a time consuming task, some approaches involved using ontology class and subclass names to generate seed examples for the learning process. Those names are used to learn contexts from the web and then those contexts are used to extract information (Kiryakov et al. 2004; Buitelaar et al. 2008). Other approaches added Hearst patterns to increase the amount of seed examples (McDowell & Cafarella 2008). For instance, considering the Bird class, useful patterns would be “birds such as X,” “birds including X,” “X and other birds,” and “X or other birds,” among others.

3.3 Related work

Relevant related work on search, particularly search applied to health information is the focus of this section. Here, relevant work related to ontology-based health information search is presented. First, recent trends in semantic search are presented; thereafter we discuss some of the trends related to health search and its specificities followed by a discussion of information extraction applied to health. The section ends with recent relevant work on OBIE.

3.3.1 Semantic search

In recent years, the interest in semantic search has increased. Even mainstream search engines such as Google or Bing are evolving to include semantics. Some systems do not assume that all, or most data, have a formal semantic annotation. One approach is expanding the user query by including synonyms and meronyms of the queried terms (Moldovan & Mihalcea 2000; Buscaldi, Rosso & Arnal 2005). Term expansion is made using the “OR” operation available in most search engines. A somewhat similar approach is followed by Kruse et al. (2005), which uses WordNet ontology and the “AND” operation of search engines to provide semantic clarification on concepts that have more than one meaning in WordNet.

Other approaches combine full text search and ontology search. ESTER (Bast et al. 2007) features an entity recognizer that assigns words or phrases to the entities of the ontology. Then, when searching for information, two basic operations are used: prefix search and join. This allows discrimination of different meanings of a concept but without logic inference. A different approach is adopted by Rocha, Schwabe & Aragao (2004). This approach involves using a regular full text search plus locating additional relevant information by using other document data such as document creator. This additional data is stored in a RDF graph, which is traversed in order to find similar concepts.

Systems that process only data with a formal semantic annotation use SPARQL queries to retrieve results (Guha & McCool 2003; Lei, Uren & Motta 2006; Esa, Taib & Thi 2010). The problems usually addressed in these cases are performance, in terms of reasoning speed, and how to rank the result set. A discussion on semantically enhanced search engines for web content discovery can be found in Kamath et al. (2013). Jindal, Bawa & Batra (2013) present a detailed review of ranking approaches for semantic search on web.

3.3.2 Health information search and exploration

In the health domain, web-available search engines are mainly targeted at retrieving information from related knowledge resources such as PubMed, the Medical Subject Headings thesaurus (MeSH) of the U.S. National Library of Medicine and the Unified Medical Language System (UMLS) of the U.S. National Library of Medicine.

CISMeF and HONselect are examples of such systems. The objective of CISMeF (Darmoni et al. 2000) is to assist health professionals and consumers in their search for electronic health information available on the Internet. CISMeF, initially only available in French, has recently improved in two ways, being currently: (1) a generic tool able to describe and index web resources and PubMed citations

or Electronic Health Records; (2) multi-lingual by allowing queries in multiple terminologies and several languages. HONselect (Boyer et al. 2001) presents medical information arranged under MeSH, offering advanced multilingual features to facilitate comprehension of web pages in languages other than those of the user.

Another health-specific information search engine is WRAPIN (Gaudinat et al. 2006). WRAPIN combines search in medical Web pages with other “hidden” online documents that are not referenced by other search engines. WRAPIN analyses a page for the most important medical terms, performing frequency analysis on MeSH terms found on the page. It identifies keywords which are then used for weighted queries to its indexes and to translate into languages other than that of the initial query. WRAPIN also allows the most important medical concepts in the document to be highlighted.

Can & Baykal (2007) designed MedicoPort, a medical search engine designed for users with no medical expertise. It is enhanced with domain knowledge obtained from UMLS in order to increase search effectiveness. MedicoPort is semantically enhanced by transforming a keyword search into a conceptual search, both for web pages and user queries.

As an example of recent work, Mendonça et al. (2012) designed and developed a proof-of-concept system for a specific group of target users and a specific domain, namely, *Neurological Diseases*. The application allows users to search for neurologic diseases, and collects a set of relevant documents with the support of ontology navigation as an auxiliary tool to redefine a query and change previous results.

Another example of recent work is that of Dragusin et al. (2013) who introduced FindZebra, a specialized rare disease search engine powered by open-source search technology. FindZebra uses freely available online medical information, but also includes specialized functionalities such as exploiting medical ontological information and UMLS medical concepts to demonstrate different ways of displaying results to medical experts. The authors concluded that specialized search engines can improve diagnostic quality without compromising the ease of use of the current and widely popular web search engines.

3.3.3 Information extraction for health

In the clinical domain, IE was initially approached with complete systems, i.e., systems including all functions required to fully analyze free-text. Examples of these large-scale projects are:

- The Linguistic String Project – Medical Language Processor of New York University
- The Specialist system (McGray et al. 1987) developed at the United States National Library of Medicine as part of UMLS project. This system includes

the Specialist Lexicon, the Semantic Network, and the UMLS Metathesaurus (USNLM 2008)

- The Medical Language Extraction and Encoding system (MedLEE) system (Friedman et al. 1995) developed at the New York Presbyterian Hospital at Columbia University. MedLEE is mainly semantically driven; it is used to extract information from clinical narrative reports, to participate in an automated decision-support system, and to allow natural language queries.

Significant resources were required to develop and implement these complete medical language processing systems. Consequently, several authors experimented over time with simpler systems that were focused on specific IE tasks and on limited numbers of different types of information to extract. Some of the areas currently benefiting from IE methods are biomedical and clinical research, clinical text mining, automatic terminology management, decision support and bio-surveillance. These narrowly focused systems demonstrated such good performance that they now constitute the majority of systems used for IE. Relevant examples of this type of system are those from the International Classification of Diseases (ICD) (Aronson et al. 2007; Crammer et al. 2007).

3.3.4 Ontology-based information extraction – OBIE

Different approaches to OBIE have been proposed and developed over the years. The approaches differ in some dimensions as follows: (1) the identification and extraction of information can be performed using probabilistic methods or explicitly defined sets of rules; (2) the types of document from which information is extracted can be unstructured, plain text, or semi-structured and structured sources; (3) the ontology can be constructed from the document's content or exist before the process started, and has the option of being updated automatically while processing documents; and (4) the kind of information extracted varies from extracting only ontological instances to extracting entire ontological classes and its associated properties.

Several IE groups focused on the development of extraction methods that use the content and predefined semantics of an ontology to perform the extraction task without human intervention and dependency on other knowledge resources (Embley et al. 1998; Maedche et al. 2002; Buitelaar & Siegel 2006; Yildiz 2007).

Considering IE for generic domains, a frequent approach is to use Wikipedia to build their knowledge base. Relevant examples of such systems were developed by Bizer et al. (2009), Suchanek, Ifrim & Weikum (2007), and Wu, Hoffmann & Weld (2008). Wikipedia structure is used to infer the semantics, and the knowledge base is populated by extracting information from pages of text and infoboxes.

Other approaches, that do not take advantage of Wikipedia structure, acquire information from generic web pages. The knowledge base structure is often inferred from pages of content and the knowledge base is populated using the same sources. Two good examples are Etzioni et al. (2004) and Yates et al. (2007).

Most approaches, whether using Wikipedia or not, use shallow linguistic analysis to detect the information to extract. Shallow analysis involves detecting text patterns and, at most, using part-of-speech information: e.g., which words are nouns, verbs, adverbs, or adjectives. The use of shallow linguistic information, however, makes it difficult to acquire information from complex sentences. A comprehensive survey of current approaches to OBIE can be found in Wimalasuriya & Dou (2010).

3.4 A general architecture for health search: handling both private and public content

Health-related information can have quite different access restrictions. Personal health-related information is confidential and has restricted access even inside health organizations. On the other hand, other health-related information, such as general information on drugs and disease characterization, is available to the general public. Not every case falls neatly into either personal health information that is confidential or more broadly into the category of general health info accessible to the public writ large. It is thus important that both well-defined cases which are either personal health-related or general-health related, and those that are not as extreme but share features of both categories, be addressed in a similar way. In this section we present a unified view of these cases (Fig. 3.1).

The main differences between the two extreme cases of personal health info versus general health pivots on (1) who are the target users of the information, i.e., if the search is made available to a general audience or only to authorized users (power users in the figure); and (2) if such a search is only possible inside the intranet of the organization holding the original source of information. In our view, these differences do not need different architectures to be handled effectively. In both cases, processing of the public and private documents can be performed by a similar pipeline, combining IE and semantic integration, and making use of ontologies. The Search and Inference Engine can also be the same. The only requirement is that the interfaces can handle access restrictions, preventing both unauthorized user access and access from the web when the information is for an organization's restricted internal use.

The proposed architecture is composed of a set of modules in which the IE component consists of a basic set of processing elements similar to the basic set

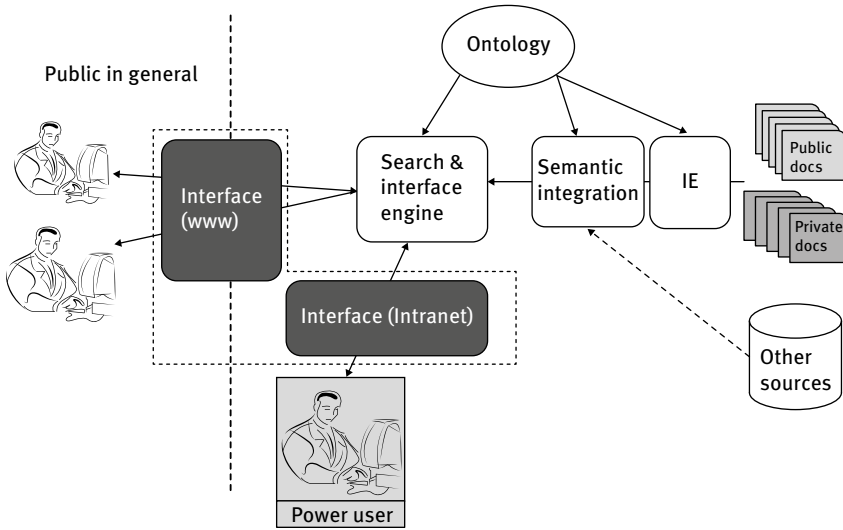


Fig. 3.1: Unified view of semantic search for health, handling both access by the general public and restricted access by authorized users inside an organization.

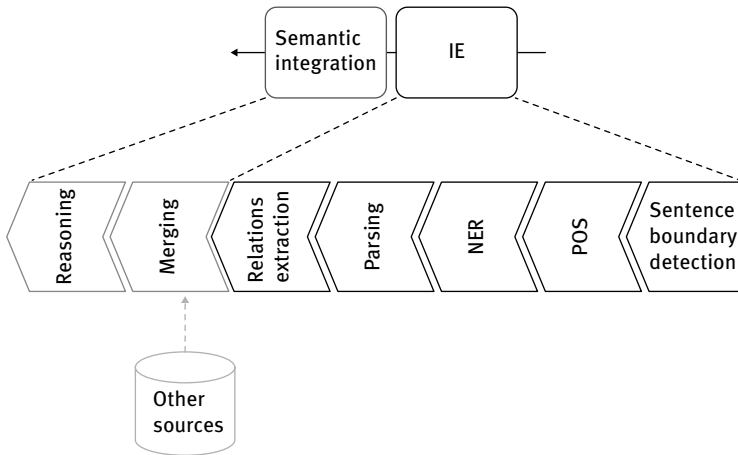


Fig. 3.2: The basic set of modules to be included in the processing pipeline to extract information from natural language sources and feeding the search engine.

illustrated in Fig. 3.2 and described by Hobbs (2002). The semantic integration module is responsible for merging the information extracted in the previous modules and reasoning. The following sections describe in detail the architecture of the two different systems used in this work.

3.5 Two semantic search systems for health

In this section we present in detail two different systems providing semantic search supported by information extraction for both private and public content. The first system targets the search inside a health institution such as a hospital, or more generally what is considered an *Intranet* search, which is a search within an organization's own internal website or group of websites. The second system in contrast targets the search outside the confines of the organization so as to enable the general public to semantically search and explore health-related information made available on the World Wide Web in Portuguese. Both of these semantic search systems were designed for Portuguese, but can be readily extended to other languages.

3.5.1 MedInX

MedInX (Ferreira, Teixeira & Cunha 2012) is a medical information eXtraction system tailored to process textual clinical discharge records in order to perform automatic and accurate mapping of free text reports onto a structured representation. MedInX is designed to be used by health professionals, and by hospital administrators and managers, as it also allows its users to search the contents of such automatically populated ontologies. MedInX uses IE technology to structure the information present in discharge reports originated by the electronic health record (EHR) system used in the region of Aveiro, Portugal in the Telematic Healthcare Network RTS® (Cunha et al. 2006). The way it works is by automatically instantiating a knowledge representation model from the free-text patient discharge letters (PDL) issued by the hospital.

During a patient's hospitalization, a large amount of data is produced in textual form as in the case of patient discharge letters. The purpose of these documents is to transfer summarized information from the hospital setting to other places, normally to the general practitioner, in order to assure continuity of patient care. MedInX addresses this type of narrative since they cover the whole inpatient period and summarize the major occurrences during that period.

The first step in development of MedInX was the creation of a corpus of authentic health records to be used in the development and evaluation of the system. This corpus was gathered through a list of hospital episodes for which a code had been assigned relative to the diagnosis of a cerebrovascular disease. If more than one cerebrovascular disease were found in the patient, additional codes pertaining to those conditions were entered in the patient's record. The corpus consists, thus, of 915 discharge letters written in Portuguese, corresponding to patients

Tab. 3.1: Number of documents, tokens and sentences in the MedInX corpus and its subsets.

Number of	Development set	Test set	Total
Documents	829	86	915
Tokens	215,730	21,788	237,518
Tokens/Document	260	253	260
Sentences	12,974	1,346	14,320
Sentences/Document	16	16	16

admitted with at least one of the several cerebrovascular disease-related codes. Table 3.1 gives statistics pertaining to the amount of documents, sentences, and tokens included in the development and test-evaluation set of MedInX.

Figure 3.3 presents a style-preserving illustration of a PDL, showing how it is possible to analyze the general content and structure of the documents. To begin with, the discharge documents have several interesting contextual features. In general, it is evident that the narratives are written from one professional to another in order to support information transfer, remind them about important medical facts, and supplement with crucial numerical data such as blood pressure and lab test results. The texts are normally intelligible and the meaning becomes evident from the context even in the presence of numerous linguistic and grammatical mistakes, word abbreviations, acronyms, signs, and other communicative features.

MedInX is a system designed for the clinical domain, which contains components for the extraction of hypertension-specific characteristics from unstructured PDLs. Its components are based on NLP principles; as such, they contain several mechanisms to read, process, and utilize external resources, such as terminologies and ontologies. These external resources represent an important part of the system by providing structured representations of the domain, clinical facts, and events that are present in the texts. The MedInX ontologies allow the assignment of domain-specific meanings to terms and use these meanings in their operations.

3.5.1.1 MedInX ontologies

In MedInX four new ontologies were created. The first two consist of two formalizations of the international classification systems that are supported by the World Health Organization (WHO): the International Classification of Diseases (ICD); and the International classification of functioning, disability and health (ICF). A drugs ontology and a conceptualization of the structure and content of the discharge reports comprise the last two of the MedInX ontologies.

<p>Motivo Internamento AVC isquémico de repetição; HTA</p> <p>História Clínica Doente de 83 anos, com antec. de HTA e depressão nervosa. Faz uso regular de fármacos que desconhece nomes. Hoje veio transferida do Hospital do Visconde de Salreu por desvio da comissura labial para esquerda. Medicada regular/ com Ogasto; Tenoretic; Micardis e Motillium. A doente refere que cerca das 09H00 teve episódio de disartria acompanhado de desvio da comissura labial a dta, sem alterações da FM ou da sensibilidade. Recorreu ao Hosp de Estarreja tendo sido encaminhada a esta Urgência por hipótese de AVC. Desde a entrada no HVS que refere melhoria progressiva das alterações da fala, sem défices neurológicos de novo. Sem queixas sugestivas de síndrome infecciosa ou de febre.</p> <p>Exame Físico COC, eufneica sem SDR. Apiretica. Chorusa. Corada e hidratada. TA- 193/68 mmHg; spO2 (AA)- 99%; PR- 60/min. AC- irregular, por ES. AP- Mv+ sem RA valorizáveis. Sem edemas dos MI's. ENS: EG- 15; sem lateralização motora; FM preservada; sem alteração da linguagem; pupilas I/R; olhos na linha média; esboço de paresia facial central a Dta.</p> <p>Terapêutica Efectuada Terapêutica efectuada: -aas 100, enoxaparina, esomeprazol, insulina sos, nitratos transdermicos, soros.</p> <p>Destino Hos. de Salreu</p> <p>Evolução Tranferida para o Hosp. de Salreu</p>	<p>Admission Reason Recurrent ischemic stroke, HTA</p> <p>Clinical History 83 years old patient, with history of HTA and clinical depression. Uses drugs regularly of which does not know names. Was transferred today from the Hospital Visconde Salreu by deviation of the left lip. Regularly medicated with Ogasto; Tenoretic; Motillium and Micardis. The patient states that at approximately 09:00 had an episode of dysarthria accompanied by deviation of the right lip without changes in MS or sensation. Appealed to the Hospital of Estarreja and was forwarded to this Urgency due to stroke suspicion. Refers progressive improvement of speech changes since entering the HVS, no neurological deficit again. No complaints suggestive of infectious syndrome or fever.</p> <p>Physical Examination COC, eupneic without RDS. Afebrile. Tearful. Healthy coloring and hydrated. BP-193 / 68 mmHg; spO2 (AA) - 99%, PR-60/min. CA-irregular, due to ES. PA-Bs + without considerable rales. No edema in the IMs. ENS: EG-15, no motor lateralization; MS preserved, without changes in language, PERRLA; eyes in the midline; outline of central facial palsy at the right.</p> <p>Therapeutics Therapeutic: -asa 100, enoxaparin, esomeprazole, sos insulin, transdermal nitrates, salines.</p> <p>Destination Salreu Hos.</p> <p>Evolution Transferred to Salreu Hosp.</p>
--	---

Fig. 3.3: A style-preserving illustration of a patient discharge letter. The original document written in portuguese is presented on the left of the figure and, on the right, its English translation.

This last ontology, in particular, was designed as an extensible knowledge model and used for storing and structuring the PDLs' entities and their relations, including temporal and modifying information. For instance, the MedInX ontology describes the fact that a Sign or Symptom is a Condition, a prescribed

Medication is a Therapeutic, a Procedure can be targeted to an Anatomical Site, and so forth. To identify the concepts in the ontology a middle-out strategy was used, i.e., first the core basic terms were identified in text and then specified and generalized as required.

3.5.1.2 MedInX system

MedInX components run within the unstructured information management architecture (UIMA) framework (Ferrucci & Lally 2004). Figure 3.4 illustrates MedInX architecture, identifying the following components:

1. **Document Reader:** a component which converts PDL files into plain text and extracts implicit meaning from the structure of the document by converting the embedded tags of the input document into annotations;
2. **General natural language processing:** components for sentence discovery, tokenization and part-of-speech tagging;
3. **REMMIX: the Named Entity Recognition component of MedInX** which concentrates on the later stages of IE, i.e., takes the linguistic objects as input and finds domain-dependent classifications and patterns among them.

REMMIX is made up of three other components:

- a. **Context Dependent annotator:** an annotator that creates annotations from one or more tokens, using regular expressions and surrounding tokens as clues;
- b. **Concept finding:** a component which extracts concepts based on specified terminologies and ontologies, and determines negation, lateralization and modifiers;
- c. **Relation extraction:** a component which extracts relations between concepts using contextual information;

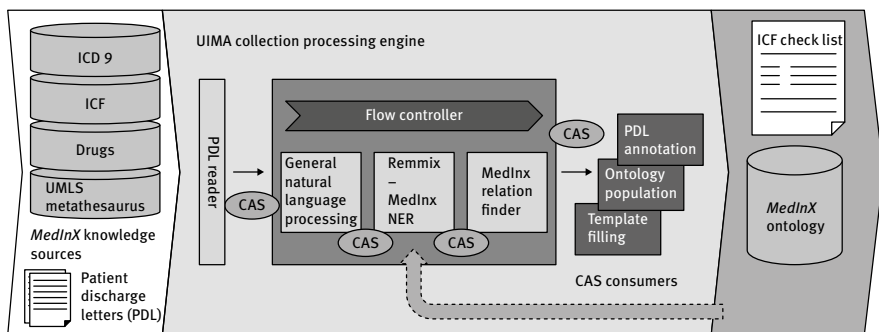


Fig. 3.4: MedInX architecture.

4. **Consumers:** responsible for ending the process and creating the desired output. The three main consumers of MedInX are:
 - a. **XML consumer:** which produces an XML file with the annotations of the previous annotators.
 - b. **Ontology population:** which populates the MedInX ontology. This component produces an OWL file with the information extracted.
 - c. **Template filling:** which outputs the knowledge extracted from the narratives to a template containing information about the patient and their health related state, with the correspondent ICF codes.

3.5.1.3 Representative results

Figure 3.5 presents MedInX evaluation interface. The extracted entities are identified by the filled boxes while the arrows represent the relations between these.

MedInX was first evaluated in 2011, in the task of extracting information from PDLs. Seven judges, belonging to different specialized areas ranging from medicine to computer science to linguistics, participated in the assessment by using the web-based evaluation interface (Fig. 3.5). To wit, the jury was made up of two computer scientists, a linguist, a radiologist, two psychologists and a physician. The 86 PDLs of the evaluation set initially selected from the MedInX corpus were automatically annotated by MedInX and made available for evaluation for

You can see and edit the result below. If the result is ok, you can press the save button to "Submit" button the annotations.

7003767
Motivo Internamento

AVC Isquémico .

História Clínica

Doente de 82 anos de idade trazida ao SU a 18/03/07 por afasia e hemiparésia à direita de início súbito, sem outras queixas acompanhantes. Trata-se de uma doente com antecedentes de patologia

Triticum 100 (R) id, Co-Diovan (R) id, Zolpidem (R) id, Apt

Exame Físico

Doente vigi, desorientada no T/E, reactiva a estímulos externos. Afasia hidratada. Apirética. Eupneica. TA = 180/90 mmHg. AC - Rítmica sem ruídos adventícios. Abdómen - globoso, mole e depressível. Ir periféricos ..

Terapêutica Efectuada

Enoxaparina, AAS, Omeprazole, Paracetamol, SOS, Captopri transdérmicos, fluidoterapia..

Entity

Category :

Type :

ICD9 :

Evaluation : Correct Incorrect Don't know

Relations

Relation	Value	Correct
HASLATERALIZATION	"direita"	<input checked="" type="checkbox"/>
HASMODIFIER	"súbito"	<input checked="" type="checkbox"/>

Enter a comment...

Cancel Save

Fig. 3.5: MedInX evaluation interface.

four months. During this four-month period a total of 30 different reports were reviewed.

The results obtained in the task of semantic classification are presented in Tab. 3.2 in terms of precision, recall, and F-measure. These values indicate a good performance of MedInX all the way around. For example, given that MedInX is tailored to the medical domain and intended to process clinical text, its proficiency with *correctly* extracting the entities and relations described in text makes it very well suited to the task. That is, only a precise system is capable of producing a correct, consistent, and concise ontology. Nonetheless, we were also concerned with the completeness of such an ontology, i.e., with the recall of the system. All in all, the results obtained indicate that MedInX performs with *both* high precision and recall, each showing an evaluation at approximately 95%. This is supported by an F-Measure, the Harmonic mean of recall and precision, whose evaluation is likewise at approximately 95%.

The clinical data included in the PDLs is a rich source of information, not only about the patient's medical condition, but also about the procedures and treatments performed in the hospital. Searching the content of the PDLs and ensuring the completeness of these documents is a process that still needs to be performed manually by expert physicians in health institutions. In order to support this process, we developed the MedInX clinical audit system. The main objective of the audit system is to help not only physicians, but hospital administrators and managers as well, to access the contents of the PDLs.

The clinical audit system uses the automatically populated MedInX ontology, which contains the structured information automatically extracted from the PDLs, and performs an automatic analysis of the content and completeness of the documents. The MedInX audit system uses the RDF query language SPARQL to query the ontology and retrieve relevant information from this resource. Several levels of information can be retrieved from this resource. An example of a developed rule, describing a complex scenario is given in Fig. 3.6. With this rule, we can find the PDLs that refer to *less than* a number of clinical Conditions and *over* a certain number of Chemical Procedures, namely, Medications and Active Substances. Both numbers used by the rule are user defined. The example of the figure uses the value 17 as the defined threshold after analysis of the most common values for these entities in the PDLs. The result of this rule allows identification of the outlier reports and suggests the need for content verification.

Tab. 3.2: Results of MedInX in the task of semantic classification.

	Precision	Recall	F-measure
Semantic classification	94.87	94.84	94.85

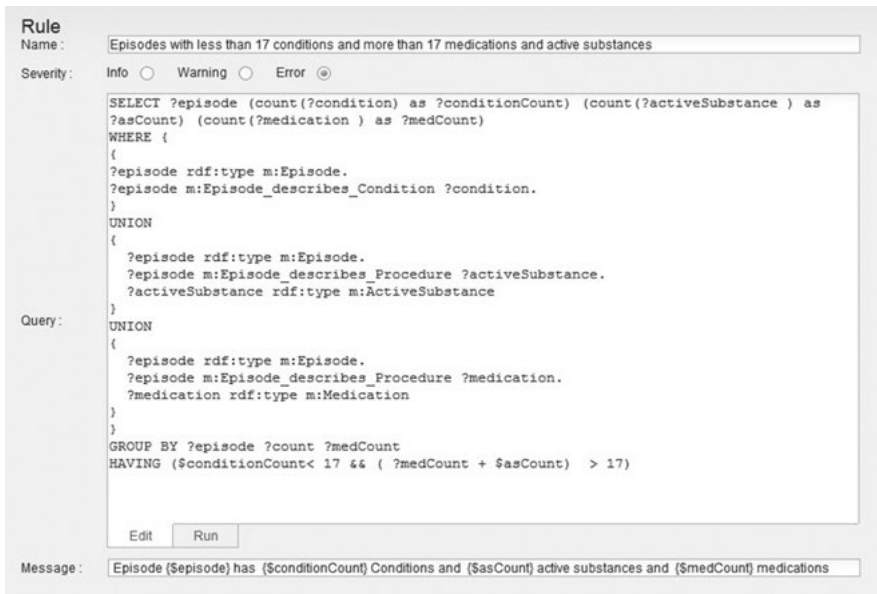


Fig. 3.6: MedInX audit system rule editor.

3.5.2 SPHInX – Semantic search of public health information in portuguese

The proof of concept system described next aims to perform semantic search on health information publicly available on the web in Portuguese.

3.5.2.1 System architecture

The SPHInX implementation is organized in four modules, respectively:

- **Natural Language Processing:** includes document content processing technologies to retrieve structured information from natural language texts. It is named NLP because it is based on technologies from the NLP area. In this part of the prototype, text is extracted from documents and enriched with the inclusion of POS tags, identification of named entities, and the construction of syntactic structures.
- **Domain Representation:** has tools for defining data semantics and associates it with samples of the NLP module output. System semantics is defined via ontology and, according to the ontology defined, it is necessary to provide examples of ontological classes and relations in sample documents. The examples are used to train semantic extraction models.
- **Semantic Extraction and Integration:** trains and applies semantic extraction models to all texts in order to obtain meaningful semantic information. It

complements the extracted information with external structured sources, e.g., geocodes and stores everything in a knowledge base conforming to the defined ontology.

- Search: information in the knowledge base can be searched and explored using natural language queries or via SPARQL.

3.5.2.2 Natural language processing

SPHInX was developed to process generic unstructured documents written in Portuguese, not only PDLs. Thus the NLP part was developed to handle the Portuguese language. The processing is organized in four sequential steps: (1) end of sentence detection; (2) Part-of-Speech (POS) tagging; (3) Named Entity Recognition (NER); and (4) syntactic parsing.

Text is extracted from documents, and then sentences are separated using the sentence boundary detector Punkt (Kiss & Strunk 2006). The sentence boundary detector step is highly relevant because all natural language processing is done in a per sentence fashion. This means that sentences define the processing context in the next NLP steps, which include algorithms able to use all the content of a sentence without using any content of the previous or following sentences.

After the split, sentences are enriched with POS tags assigned by TreeTagger (Schmid 1994): noun, verb, adjective, etc. TreeTagger was trained with a European Portuguese lexicon in order to be integrated in the system. Its outputs contain the word form followed by the assigned POS tag and the word lemma.

Named entities are discovered and classified by REMBRANDT (Cardoso 2012). Words belonging to a named entity are grouped using underscores. For instance, the names of the person John Stewart Smith become the single token John_Stewart_Smith. Then, sentences are analyzed to determine their grammatical structure. This is done by MaltParser (Hall et al. 2007) and the result is a planar graph encoding the dependency relations among the words of each sentence.

3.5.2.3 Semantic extraction models

SPHInX creates one semantic extraction model for each ontology class and ontology relation. A model is a set of syntactic structure examples and counter examples that were found to encode the meaning represented by the model. It also contains a statistical classifier that measures the similarity between a given structure and the model's internal examples. The model is said to have positively evaluated a sentence fragment if the similarity is higher than a given threshold.

The algorithm for creating semantic extraction models was inspired in two studies. The first is about extracting instances of binary relations using deep syntactic analysis. Suchanek, Ifrim & Weikum (2006) extracted one-to-one

and many-to-one relations such as place and date of birth. They used custom-built decision functions to detect facts for each relation, and a set of statistical classifiers to decide if new patterns are similar to the learned facts. In our proof-of-concept prototype, this work was extended to include the extraction of one-to-many and many-to-many relations. The proof-of-concept prototype also implements a general purpose decision function based on the annotated examples instead of a custom-built function for each relation.

The second work is about improving entity and relation extraction when the process is learned from a small number of labeled examples, using linguistic information and ontological properties (Carlson et al. 2009). Improvements are made using class and relation hierarchy, information about disjunctions, and confidence scores of facts. This information is used to bootstrap more examples thereby generating more data to train statistical classifiers. For instance, when the system is confident about a fact, such as when it was annotated by a person, this fact is used as an instance of the annotated class and/or relation. This fact can also be used as a counter-example of all classes/relations disjoint with the annotated class/relation, and as an instance of super-class/super-relation. Moreover, facts discovered by the system with a high confidence score can be promoted to examples and included in a new round of training. In the proof-of-concept prototype, this creation of more examples is not active by default as it can lead to data over-fitting and should therefore be used carefully.

For the first version of SPHInX, the ontology about neurological diseases used in Mendonça et al. (2012) was adopted. The semantic extraction models were trained with a set of six manually annotated documents, of around fifty pages each, by a person familiar with the ontology but not related to the prototype development. The annotations were related to neurological diseases and respective symptoms, risk factors, treatments and related drugs.

3.5.2.4 Semantic extraction and integration

All sentence graphs are evaluated by the classifiers of all semantic models, and are collected in the case of forming a triple. A sentence fragment forms a *triple* if it is positively evaluated by two class models, one for subject and the other for object, along with one relation model binding the subject and object (Rodrigues, Dias & Teixeira 2011). Missing information according to the ontology is searched in external structured information sources. For instance, unknown locations of entities with a fixed place (such as streets, organizations' headquarters, and some events) are queried using Google Maps API.

All collected triples are tentatively added to the knowledge base and their coherence is verified by a semantic reasoner. In SPHInX, reasoning is performed by an open-source reasoner for OWL-DL named Pellet (Sirin & Parsia 2004). All triples not coherent with the rest of the knowledge base are discarded, and a warning is issued. The remaining triples become part of the knowledge base.

3.5.2.5 Search and exploration

The search and exploration part of the system will be explained based on an illustrative example of use. The data for this example was obtained by having the system process fourteen previously unseen documents using the semantic extraction models trained earlier, plus the data already on the ontology at the time. Then, the same person that annotated the training documents was asked to suggest a few possible questions in Portuguese. Those questions were submitted to the system and one of them was selected for the example.

The interface, based on NLP-Reduce (Kaufmann, Bernstein & Fischer 2007), accepts natural language questions and generates SPARQL queries that are passed to a SPARQL engine. The system allows the user to enter a sentence, such as “memory loss is a symptom of what diseases?”

First, the question is transformed by removing all stop words and punctuation marks. The remaining words are stemmed and passed to a query generator that will use them to produce a SPARQL query in four steps:

1. Search for triples that contain one or more words of the query in the object property label. Triples are ranked according to the amount of words included in the label.
2. Search for properties that can be joined with the triples found in step 1. Thus, properties are searched using domain and range information of triples from step 1 along with the remaining query words. In the case of query words producing triples based on alternative object properties, the triples favored are those with the highest score from step 1. The triple set of this step is combined with the set of step 1, according to the ontology rules.
3. Search for data type property values that match the query words not matched in steps 1 and 2. Triples found are once again ranked considering the amount of words included in the property values. All triples found respecting the domain and range restrictions of the set created in step 2 are added to it.
4. When there are no more query words left, the SPARQL query is generated to join the retrieved triples that achieved the highest scores in steps 1 to 3. Semantically equivalent duplicates are removed and the query is ready to be passed to a SPARQL endpoint.

So a question like “a perda de memória é um sintoma de que doenças?” which is roughly the Portuguese equivalent to “memory loss is a symptom of what diseases?” originates the SPARQL code presented in Fig. 3.7.

The output of the query is a table containing the variables of the SPARQL query. Depending on the type of information asked, the data on the table can be plain text as in case of data type property, presented in Tab. 3.3, or links to other ontological entities as in the case of object properties. In the case of the latter, it is then possible to navigate through the ontology by following those links.

Another way to output results is by presenting the graph of ontological concepts involved in the query. Figure 3.8 depicts the graph of the ontological elements used to compute the answer to the example query. As can be seen, there are more concepts involved than the ones included in the output table. The concepts involved in queries and not presented in the output table are typically the ones used to compute logical inferences.

Presenting the results in a graphic format allows users to navigate the information stored in the knowledge base while keeping a good overview of the ontology and how different concepts relate to each other.

SPARQL code	Explanation
<pre>prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> prefix owl: <http://www.w3.org/2002/07/owl#> prefix MedInX: <http://www.MedInX.com/MedInX.owl#> select distinct * WHERE { ?NamedIndividual MedInX:hasSinalSintoma ?Sinal_Sintoma . ?Sinal_Sintoma rdfs:label ?Sinal_Sintoma_label . FILTER(REGEX(?Sinal_Sintoma_label,'memoria','i')). ?NamedIndividual rdfs:label ?NamedIndividual_label . ?Sinal_Sintoma rdf:type MedInX:Sinal_Sintoma . ?NamedIndividual rdf:type owl:NamedIndividual }</pre>	<p>Get instances of relation hasSinalSintoma</p> <p>Get labels of the relation objects</p> <p>... and accept those including word “memoria”</p> <p>Get labels of the relation subjects</p> <p>Relation objects have the type Sinal_Sintoma</p> <p>... and subjects have the type NamedIndividual</p>

Fig. 3.7: SPARQL code generated by the natural language interface.

Tab. 3.3: Result set for the query “a perda de memória é um sintoma de que doenças?” (“memory loss is a symptom of what diseases?”).

NamedIndividual	Sinal_Sintoma	Sinal_Sintoma_label	NamedIndividual_label
“Doença de Parkinson”	“Perda de memória”	“Perda de memória”@pt	“Doença de Parkinson”@pt
“Doença de Huntington”	“Perda de memória”	“Perda de memória”@pt	“Doença de Huntington”@pt
“Esclerose Multipla”	“Perda de memória”	“Perda de memória”@pt	“Esclerose Multipla”@pt
“Demencia de tipo Alzheimer”	“Perda de memória”	“Perda de memória”@pt	“Demencia de tipo Alzheimer”@pt

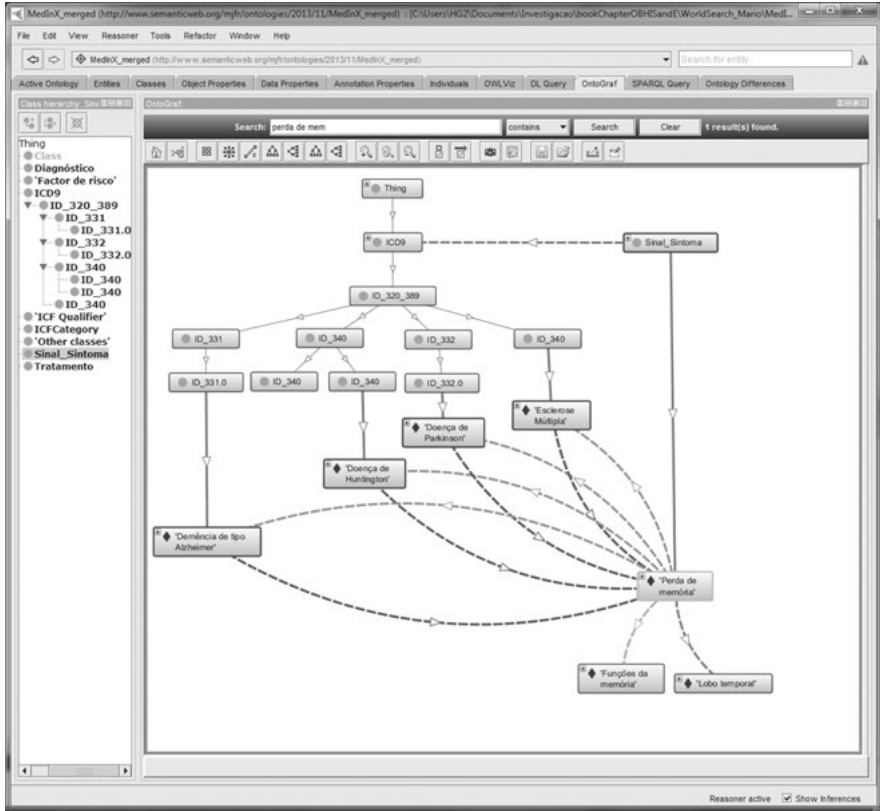


Fig. 3.8: Graph of the ontology concepts involved in the query example.

3.6 Conclusion

Taking into consideration the increasing need for semantic search of health information available originally in natural language, in this chapter a general architecture predicated on ontology-based information extraction to feed a search engine is proposed and instantiated in two systems. The first system, MedInX, allows semantic search of the information regarding a hospital’s discharge letters, and can be generalized to the vast information in natural language stored in internal web-based hospital information systems. The second system, SPHInX, currently at an early stage of development, is capable of extracting information from public documents in Portuguese. For both systems, we present information on its architecture and components, and show via demonstration how these systems work.

We envision future developments of both systems that would address a much broader area, as both systems that we presented here only address a limited

medical domain. Another equally important goal pivots on the improvement of the interaction of the user with these systems, making search and exploration a natural experience for professionals and laity alike.

Acknowledgments

This work was partially supported by World Search, a QREN project (QREN 11495) co-funded by COMPETE and FEDER, and the Portuguese Foundation for Science and Technology PhD grant SFRH/BD/27301/2006 to Lílíana Ferreira. The authors also acknowledge the support from IEETA Research Unit, FCOMP-01-0124-FEDER-022682 (FCT-Pest C/EEI/UIO127/2011).

References

- Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P. & Vilain, M. (1995) MITRE: description of the Alembic system used for MUC-6. *Proceedings of the 6th conference on Message understanding* (pp. 141–155). New York, NY: Association for Computational Linguistics.
- Abraham, J. & Reddy, M. (2007) Quality of healthcare websites: A comparison of a general-purpose vs. domain-specific search engine. *AMIA Annual Symposium Proceedings*, 858.
- Aronson, A., Bodenreider, O., Demmer-Fushman, D., Fung, K., Lee, V. & Mork, J. (2007) From indexing the biomedical literature to coding clinical text. *BioNLP '07: Proceedings of the Workshop on BioNLP 2007*, (pp. 105–112).
- Bast, H., Chitea, A., Suchanek, F. & Weber, I. (2007) ESTER: Efficient Search on Text, Entities, and Relations. *Proceedings of the 30th ACM SIGIR*, (pp. 679–686).
- Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F. & Secret, A. (1994) 'The World-Wide Web', *Commun ACM*, 37:76–82.
- Berners-Lee, T. & Fischetti, M. (1999) *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: Harper Collins.
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. & Hellmann, S. (2009) DBpedia – A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web – The Web of Data*, 7, 154–165.
- Boyer, C., Baujard, V., Griesser, V. & Scherrer, J. R. (2001) 'HONselect: a multilingual and intelligent search tool integrating heterogeneous web resources', *Int J Med Inform*, 64:253–258.
- Buitelaar, P. & Siegel, M. (2006) Ontology-based Information Extraction with SOBA. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, (pp. 2321–2324).
- Buscaldi, D., Rosso, P. & Arnal, E. S. (2005) A wordnet-based query expansion method for geographical information retrieval. *Working Notes for the CLEF Workshop*.
- Can, A. B. & Baykal, N. (2007) 'MedicoPort: a medical search engine for all', *Comput Meth Prog Biomed*, 86:73–86.

- Cardoso, N. (2012) 'Rembrandt – a named-entity recognition framework'. *LREC*, (pp. 1240–1243).
- Carlson, A., Betteridge, J., Hruschka, E. R. & Mitchell, T. M. (2009) Coupling Semi-Supervised Learning of Categories and Relations. *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (pp. 1–9). Association for Computational Linguistics.
- Cowie, J. & Lehnert, W. (1996) 'Information extraction', *Communications of the ACM*, 39:80–91.
- Cramer, K., Dredze, M., Ganchev, K., Talukdar, P. P. & Carroll, S. (2007) Automatic code assignment to medical text. *BioNLP '07: Proceedings of the Workshop on BioNLP 2007* (pp. 129–136). Association for Computational Linguistics.
- Cunha, J. P., Cruz, I., Oliveira, I., Pereira, A. S., Costa, C. T., Oliveira, A. M. & Pereira, A. (2006) The RTS project: Promoting secure and effective clinical. *eHealth 2006 High Level Conference*, (pp. 1–10).
- Darmoni, S. J., Leroy, J. P., Baudic, F., Douyère, M. A. & Thirion, B. (2000) 'CISMeF: a structured health resource guide', *Met Inform Med*, 30–35.
- Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jorgensen, H. L., Cox, I. J., Hansen, L. K., Ingwersen, P. & Winther, O. (2013) 'FindZebra: a search engine for rare diseases', *Int J Med Inform*, 82:528–538.
- Embley, D. W., Campbell, D. M., Smith, R. D. & Liddle, S. W. (1998) Ontology-based extraction and structuring of information from data-rich unstructured documents. *Proceedings of the seventh international conference on Information and knowledge management* (pp. 52–59). ACM.
- Esa, A. M., Taib, S. M. & Thi, H. N. (2010) Prototype of semantic search engine using ontology. *Open Systems (ICOS), 2010 IEEE Conference on* (pp. 109–114). IEEE.
- Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. & Yates, A. (2004) Web-Scale Information Extraction in KnowItAll (Preliminary Results). *WWW '04 – Proceedings of the 13th International World Wide Web Conference* (pp. 100–110). New York, NY, USA: Association for Computational Linguistics.
- Ferreira, L., Teixeira, A. & Cunha, J. P. (2012) *Medical Information Extraction – Information Extraction from Portuguese Hospital Discharge Letters*. Lambert Academic Publishing.
- Ferrucci, D. & Lally, A. (2004) 'UIMA an architectural approach to un-structured information', *Nat Lang Eng*, 10:327–348.
- Friedman, C., Johnson, S. B., Forman, B. & Starren, J. (1995) 'Architectural requirements for a multipurpose natural language processor in the clinical environment'. *Proc Annu Symp Comput Appl Med Care*, (pp. 347–351).
- Gaizauskas, R. & Wilks, Y. (1998) 'Information extraction: beyond document retrieval', *J Doc*, 54:70–105.
- Gaudinat, A., Ruch, P., Joubert, M., Uziel, P., Strauss, A., Thonnet, M., Baud, R., Spahn, S., Weber, P., Bonal, J., Boyer, C., Fieschi, M. & Geissbuhler, A. (2006) 'Health search engine with e-document analysis for reliable search results', *Int J Med Inform*, 75:73–85.
- Gruber, T. R. (1993) 'A translation approach to portable ontology specifications', *Knowledge Acquisition*, 5:199–220.
- Guarino, N. (1998) Formal Ontology in Information Systems. *FIOS'98 – Proceedings of the First International Conference on Formal Ontology in Information Systems* (pp. 3–15). IOS Press.
- Guha, R. & McCool, R. (2003) 'TAP: a Semantic Web platform', *Computer Networks*, 557–577.
- Guha, R., McCool, R. & Miller, E. (2003) Semantic search. *Proceedings of the 12th international conference on World Wide Web* (pp. 700–709). ACM.

- Hahn, U., Romacker, M. & Schulz, S. (2002) 'Creating Knowledge Repositories From Biomedical Reports: The MEDSYNDIKATE Text Mining System'. *Pac Symp Biocomput*, (pp. 338–349).
- Hall, J., Nilsson, J., Nivre, J., Eryigit, G., Megyesi, B., Nilsson, M. & Saers, M. (2007) *Single Malt or Blended? A Study in Multilingual Parser Optimization*. (pp. 933–939). Association for Computational Linguistics.
- Hobbs, J. R. (2002) 'Information extraction from biomedical text', *J Biomed Inform*, 35:260–264.
- Jindal, V., Bawa, S. & Batra, S. (2014) 'A review of ranking approaches for semantic search on Web'. *Inf Process Manage*, 50(2): 416–425.
- Kamath, S. S., Piraviperumal, D., Meena, G., Karkidholi, S. & Kumar, K. (2013) A semantic search engine for answering domain specific user queries. *Communications and Signal Processing (ICCSPP), 2013 International Conference on* (pp. 1097–1101). IEEE.
- Kaufmann, E., Bernstein, A. & Fischer, L. (2007) NLP-Reduce: A "naive" but Domain-independent Natural Language Interface for Querying Ontologies. *ESCW'07 – Proceedings of the 6th International Semantic Web Conference*.
- Kiss, T. & Strunk, J. (2006) 'Unsupervised multilingual sentence boundary detection', *Compu Linguist*, 32:485–525.
- Kruse, P. M., Naujoks, A., Rosner, D. & Kunze, M. (2005) Clever search: A wordnet based wrapper for internet search engines. *arXiv preprint cs/0501086*.
- Lee, L. (2004) "I'm sorry Dave, I'm afraid I can't do that": Linguistics, Statistics, and Natural Language Processing circa 2001. In C. O. Board, *Computer Science: Reflections on the Field, Reflections from the Field* (pp. 111–118). Washington DC: The National Academies Press.
- Lei, Y., Uren, V. & Motta, E. (2006) Semsearch: A search engine for the semantic web. *Proceedings of the 15th International Conference on Managing Knowledge in a World of Networks* Berlin, Heidelberg (pp. 238–245). Berlin, Heidelberg: Springer-Verlag.
- Maedche, A., Maedche, E., Neumann, G. & Staab, S. (2003) *Bootstrapping an Ontology-based Information Extraction System*, pp. 345–359. Heidelberg, Germany: Physica-Verlag GmbH.
- Màrquez, L., Carreras, X., Litkowski, K. C. & Stevenson, S. (2008) 'Semantic role labeling: an introduction to the special issue', *Comput Linguist*, 34:145–159.
- McGray, A. T., Sponsler, J. L., Brylawski, B. & Browne, A. (1987) 'The role of lexical knowledge in biomedical text understanding'. *SCAMC*, (pp. 103–107).
- McNaught, J. & Black, W. (2006) Information extraction. In Ananiadou, S. & McNaught, J. *Text Mining for Biology and Biomedicine* (pp. 143–178). Norwood: Artech House.
- Mendonça, R., Rosa, A. F., Oliveira, J. L. & Teixeira, A. J. (2012) Towards Ontology Based Health Information Search in Portuguese – A case study in Neurologic Diseases. *CISTI'2012 – 7th Iberian Conference on Information Systems and Technologies*.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990) 'Introduction to wordnet: An on-line lexical database', *Int J Lexico*, 235–244.
- Moldovan, D. I. & Mihalcea, R. (2000) 'Using wordnet and lexical operators to improve internet searches', *Internet Comput, IEEE*, 4:34–43.
- Pakhomov, S. A. (2005) High throughput modularized NLP system for clinical text. *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions* (pp. 25–28). New York, NY: Association for Computational Linguistics.
- Rocha, C., Schwabe, D. & Aragao, M. P. (2004) A hybrid approach for searching in the semantic web. *Proceedings of the 13th international conference on World Wide Web* (pp. 374–383). ACM.

- Rodrigues, M., Dias, G. P. & Teixeira, A. (2011) Ontology Driven Knowledge Extraction System with Application in e-Government. *Proc. of the 15th APIA Conference*, (pp. 760–774).
- Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*.
- Sirin, E. & Parsia, B. (2004) Pellet: An OWL DL Reasoner, In Haarslev, V. and Möller, R. (eds), International Workshop on Description Logics (DL'04), pp. 212–213. British Columbia, Canada: Whistler.
- Suchanek, F. M., Ifrim, G. & Weikum, G. (2006) LEILA: Learning to Extract Information by Linguistic Analysis. *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 18–25). New York, NY: Association for Computational Linguistics. Sydney, Australia
- Suchanek, F. M., Kasneci, G. & Weikum, G. (2007) YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. *WWW '07 – Proceedings of the 16th International World Wide Web Conference* (pp. 697–706). New York, NY: Association for Computational Linguistics.
- USNLM. (2008) UMLS Knowledge Sources. United States National Library of Medicine.
- Wang, L., Wang, J., Wang, M., Li, Y., Liang, Y. & Xu, D. (2012) 'Using internet search engines to obtain medical information: a comparative study'. *J Med Internet Res*, 14.
- Wang, C., Xiong, M., Zhou, Q. & Yu, Y. (2007) PANTO: A Portable Natural Language Interface to Ontologies. *ESWC2007 – Proceedings of the 4th European Semantic Web Conference* (pp. 473–487). Berlin/Heidelberg: Springer.
- Wimalasuriya, D. C. & Dou, D. (2010) 'Ontology-based information extraction: An introduction and a survey of current approaches', *J Inform Sci*, 36:306–323.
- World Health Organization. (n.d.) *International Classification of Diseases*. Accessed on 01/27/2014, <http://www.who.int/classifications/icd/en/>
- Wu, F., Hoffmann, R. & Weld, D. S. (2008) Information Extraction from Wikipedia: Moving Down the Long Tail. *KDD '08 – Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 731–739). New York, NY: Association for Computational Linguistics.
- Yates, A., Banko, M., Broadhead, M., Cafarella, M. J., Etzioni, O. & Soderland, S. (2007) TextRunner: Open Information Extraction on the Web. *NAACL-HLT (Demonstrations) – Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 25–26). Morristown, NJ: Association for Computational Linguistics.
- Yildiz, B. (2007) Ontology-driven Information Extraction. *PhD Thesis*. Vienna University of Technology.

Part II Machine Learning Techniques for Mining Medical Search Queries and Health-Related Social Media Posts and Tweets

Jedsada Chartree, Angel Bravo-Salgado, Tamara Jimenez and Armin R. Mikler

4 Predicting dengue incidence in Thailand from online search queries that include weather and climatic variables

Abstract: This chapter presents machine learning techniques to help public health agencies mitigate vector borne disease, in particular dengue outbreaks. The methods presented in this study will predict the number of dengue cases so that public health authorities may devise adequate interventions to address dengue outbreaks. Search queries from digital sources are used to forecast the number of dengue cases prior to officially reported cases. This is achieved by processing query terms related to vector-borne dengue disease. Climate has been correlated to the vector's dynamics; hence, query terms related to weather are utilized for the forecasting of dengue cases.

4.1 Introduction

In recent years the significance of the terms used in Internet searches has become increasingly evident. In particular, web search data can be used as trend indicators in a variety of fields. Making use of its direct access to mining search queries, Google offers the web service Google Trends, which presents frequency and location of search terms in different formats (Choi & Varian 2012). Such trends include the categories shopping, arts and companies. Third party providers make use of information obtained from search engines to target specific areas of interest, such as most desired travel locations as defined by Google searches (Shankman 2012). Information about user characteristic has been mined by analyzing sequences of searches and search modifications (Jansen et al. 2000), as well as the use of browsing history to provide personalized search results for different users despite the use of the same search terms (Sugiyama et al. 2004). Similarly, social network analysis techniques have been applied to mine the Web, blogs and online forums to predict long-term trends on the popularity of concepts. Such trends include the popularity of brands, outcomes of political elections, and the winners of movie awards (Gloor et al. 2009). Correlation between health data and information mined from the Internet has been shown in different studies. As an example, Google Flu Trends uses this information to estimate flu activity in the United

States. Due to delays in reporting and underreporting in the healthcare system, these search trends may instead serve as early predictors of outbreaks. Regional outbreaks can be detected as early as 7–10 days before surveillance systems by the Centers for Disease Control and Prevention (CDC) (Carneiro & Mylonakis 2009).

4.1.1 Dengue disease in the world

The CDC's division of vector-borne disease (DVBD) recognizes dengue as one of the vector-borne diseases in its priority list (CDC 2013). Dengue, specifically, is a viral disease that is transmitted via mosquito bites to humans. Dengue, as a reemerging disease, has gained the attention of international health agencies, such as the World Health Organization (WHO) and the Pediatric Dengue Vaccine Initiative (PDVI). The PDVI was founded in 2001 to advance the development of pediatric dengue vaccines for use in developing countries. Based on their collaboration with the WHO, governments, industry, and the scientific community, the PDVI has been actively involved in the development of improved policies and guidelines for clinical evaluations and vaccine testing. Now their work is continued by the Dengue Vaccine Initiative (DVI), a consortium of four organisations: The International Vaccine Institute, the World Health Organization, the International Vaccine Access Center, and The Sabin Vaccine Institute (The Dengue Vaccine Initiative 2014).

The WHO and the PDVI have estimated between 2.5 to 3.6 billion people to be at risk of contracting dengue. In 2010, about 50 million dengue cases resulting in 22,000 deaths were reported worldwide (see Fig. 4.1) (WHO 2010). In the same year, the US territory of Puerto Rico experienced the longest and largest incidence of dengue cases reported since the 1960s (Sharp et al. 2013).

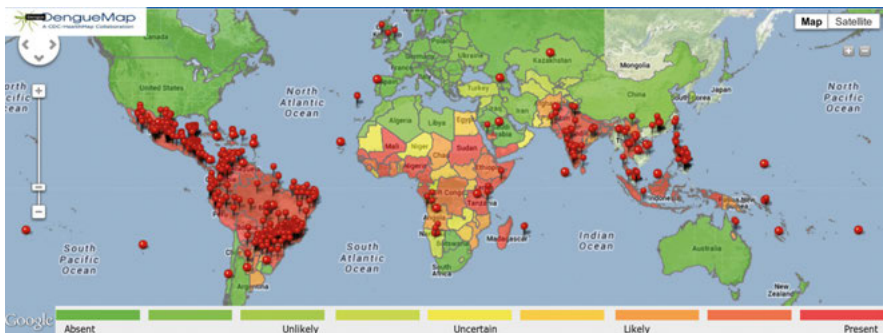


Fig. 4.1: Map of human dengue cases in 2013. Generated and adapted from (healthmap.org 2013).

Studies have shown that variation in temperature acts as an important confounder affecting metabolic processes and the reproduction of the pathogen within a vector. In addition, variation in temperature influences the number of laid eggs; thereby, modifying the dynamics of the vector population; consequently, the incidence of dengue in the human population. Peak numbers of human dengue cases during hot-dry and rainy seasons have been reported (Altizer et al. 2006). Moreover, it has been hypothesized that *transovarial*¹ transmission and infected eggs during the diapause² state may be a contributor of the survival of the virus during inter-endemic periods (Xiao-Xia et al. 2004).

The use of analytic tools such as mathematical and computational models, permit both quantitative and qualitative analysis of virtual contagion scenarios and the forecasting of future outbreaks. Computational disease models allow public health experts to investigate the impact of each mode of transmission, both horizontal and vertical transmission. The ubiquitous access to the Internet and the accessibility of information stored in data centers have given the scientific community a unique opportunity to apply modern statistical techniques to the study of disease dynamics. Making use of these scientific tools, public health authorities can anticipate imminent dengue outbreaks, establish adequate preparedness plans, and develop effective policies that provide rapid interventions and optimal utilization of public health resources. The systematic study of virtual disease scenarios and forecasting will allow health authorities to develop the necessary skills and capabilities today to cope with real sanitary emergencies of tomorrow.

The transmission risk of dengue is strongly dependent on the ambient temperature of the geographic region. Temperature will determine the rate at which mosquitoes develop from egg to adult and therefore the number of vectors. Further, the extrinsic incubation period (EIP), which describes the time until an infected vector becomes infectious is temperature dependent. Therefore, it is imperative for predictive models to include data about climate and climate change.

4.2 Epidemiology of dengue disease

There are two species of vectors that are a natural reservoir of dengue disease; these are *Aedes aegypti* and *Aedes albopictus*. The causative agent of dengue is in the category of arboviruses. Four serotype³ strains of the virus DEN-1, -2, -3, and -4,

1 Transovarial: dengue transmission from female mosquito into laid eggs.

2 Diapause: phase of suspended development during unfavorable climate conditions.

3 Serotype: it is a distinguishable ribonucleic acid strain of the virus.

are known to induce dengue fever (DF), dengue hemorrhagic fever (DHF), and life-threatening dengue shock syndrome (DSS). The four dengue serotypes are transmitted and disseminated into the human population through the bites of infected host mosquitoes (Anderson & May 2002).

Once a person acquires dengue, the onset of symptoms begins approximately 4–7 days after the mosquito bite. This time period is known as the latent period. The duration of symptoms typically lasts approximately 3–10 days. It is not unusual that infected people show no symptoms. A large number of dengue virions must be present in the blood stream of the infected person in order to transmit the disease back to the vectors (CDC 2010). Human infection caused by one of the serotypes leads to specific antibodies that provide long lasting immunity. After recovering from a dengue infection, a period of approximately 12 weeks (84 days), immunity is acquired for the other serotypes (Krause 1997).

While feeding on human infected blood, a mosquito digests the dengue virions that will reproduce and infect the vectors cells. Once the virions have reached the salivary glands of the mosquito, it is said to be an infectious mosquito, which will remain in this condition for the rest of its adult life. The elapsed time between digestion and salivary glands infection is known as the extrinsic incubation period (EIP). It is estimated that the EIP can last approximately 8–12 days (CDC 2010). The transmission of dengue disease that occurs between the vector and human population and vice versa is known as horizontal transmission. The annual reemergence of dengue outbreaks varies according with the vector abundance. Moreover, the abundance or the dynamics of the vector population is determined, among other factors, by weather change.

4.2.1 Temperature change and the ecology of *A. aegypti*

Temperature variation affects the phases of the *A. aegypti* life cycle. The mosquito life cycle encompasses four metamorphic phases: egg, larva, pupa, and adulthood. Effects in the life cycle of the vector are the shortening or stretching of embryo maturation; the augmentation or decreasing of the viability; and the prolonging or reduction of the life span of the mosquitoes. The influence of temperature on two of these biological phases in eggs, larvae, and pupae are depicted in Fig. 4.2.

Temperature variation not only influences the completion time of biological processes but also establishes the duration of the virus's extrinsic incubation

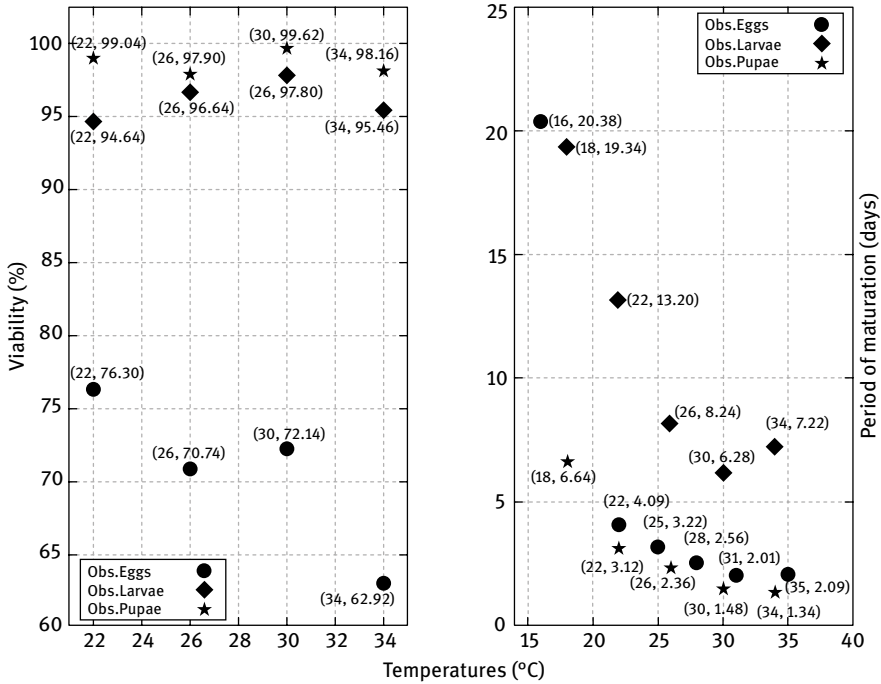


Fig. 4.2: Average viability and maturation time of the egg, larva, and pupa phases of the *A. Aegypti* mosquito (Beserra et al. 2006; Farnesi et al. 2009).

period (EIP) within the host (known as latent period). The effect of temperature in the efficacy of dengue transmission from *A. aegypti* into susceptible primate hosts have been reported by Burke et al. (Watts et al. 1987).

Unfavorable weather conditions affect the reproduction of the dengue virus within vectors. At temperatures below 24°C, the EIP stretches beyond the life span of mosquitoes. Since the end of the EIP determines the start of the vector’s infectious period, it becomes apparent that it is unlikely that transmission of the virus through mosquito biting could take place, giving no opportunity to the continuation of an endemic process.

There exists another form of dengue transmission, the transovarial transmission, which seems to contribute to the perennial and recurrent annual outbreaks. The transovarial transmission is the transmission of the dengue virus from an infected female vector to its offspring. It is known that infected eggs, under unfavorable conditions, enter into a diapause state, maintaining a dormant generation of future infected mature vectors (Xiao-Xia et al. 2004; Bennet & Joshi 2008).

In Fig. 4.3, the change in duration of the extrinsic incubation period by means of change in temperature and a flow diagram describing the transovarial transmission of dengue from a mature vector into the immature phases of the vector are presented.

To this point, we discussed the relationship among temperature, dengue virus, and disease-host *A. aegypti*. Thus far, the description of these relationships was centered at the organism level; however, another type of relationship exists between the host vector *A. aegypti* and the dengue virus which can be traced to the level of gene expression (Shuzhen et al. 2012). Knowing the importance of these relationships, each one can be taken into consideration in the design of detailed computational tools to anticipate the trends of seasonal dengue outbreaks. Detailed computational tools necessitate specific data to be available to design such methods. The ubiquity of the Internet and online data

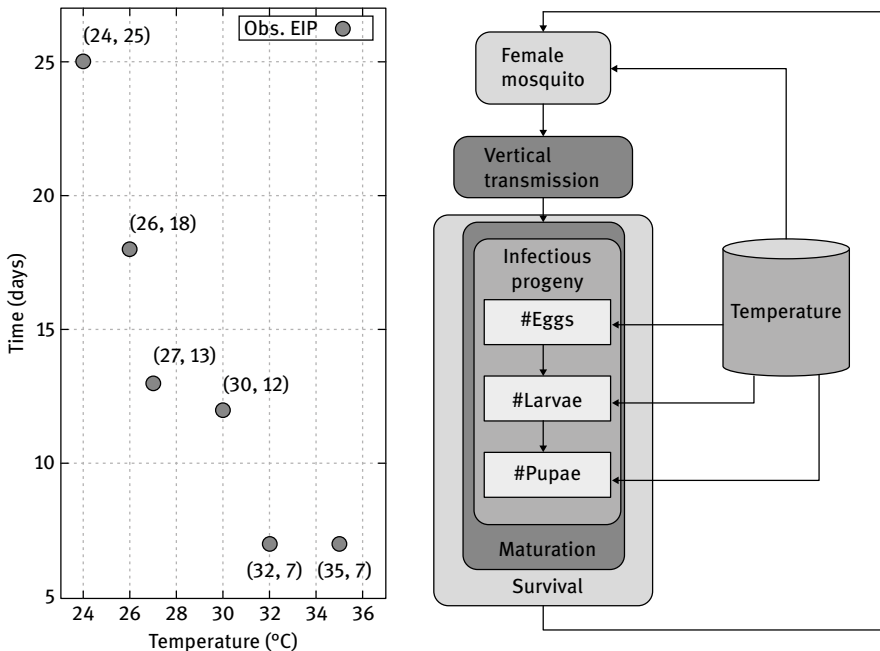


Fig. 4.3: Distribution of dengue virus at different temperatures in head, thorax-abdomen, and salivary glands of the *A. aegypti* have been reported in (Watts et al. 1987). The graph on the left displays the observed time when the presence of the virus was detected in the salivary glands. The graph on the right exemplifies both the influence of temperature in the phases of the mosquito life cycle and vertical transmission of dengue.

sources have become valuable assets for researchers and scientists in the field of public health. Available cyberinfrastructure where online activity, statuses, and queries of millions of people are stored every day has encouraged the advancement and development of methods to predict seasonal outbreaks. In previous work, the use of cyberinfrastructure and computational methods have allowed researchers to forecast yearly trends of infectious disease (Corley et al. 2009). These results have motivated the use of such cyberinfrastructure to design and implement predictive models to the prediction of vector-borne diseases. Recognizing the existing relationships among weather, host, and disease, it is the intent of the following section to introduce available methodologies that can serve as the guiding oracle for the public health experts in the control of vector-borne diseases.

4.3 Using online data to forecast incidence of dengue

This section demonstrates the effort of monitoring dengue outbreaks or forecasting the number of dengue cases to support public health authorities in the planning to mitigate dengue epidemics. This study utilized data from the Internet to predict the number of dengue cases ahead of official reported cases. Several machine learning algorithms were applied to build predictive models. This research aims to confirm that online data are useful for disease surveillance and compare the performance of predictive models.

4.3.1 Background and related work

Dengue is currently the most serious vector-borne disease globally (Guha-Sapir & Schimmer 2005; Racloz et al. 2012). The outbreaks of dengue have affected a large number of people throughout the world, especially in tropical and sub-tropical nations (Map 2012; WHO 2012). Mitigation of this particular vector-borne disease requires a tremendous amount of monetary resources every year, especially in hospitals, where the number of patients grows rapidly during an epidemic. To cope with such outbreaks efficiently, governments require experts, equipment, medicine, and response strategies to assist patients adequately and to reduce the mosquito population.

There are different challenges during a vector-borne epidemic, which include the lack of vaccine to prevent the infection; severe flu-like illness in children with weak immune systems; the lack of solid plans to control the disease's outbreak;

and the lack of sufficient specialists (Palaniappan & Awang 2008; Potts et al. 2010; Madoff et al. 2011). Governmental budget constraints, and administration inefficiencies are other factors that will result in ineffective response plans. Furthermore, a short incubation period of 4–7 days leads to a quick spread of dengue, which mostly occurs in developing countries (WHO 2012) and often affects a large number of people.

Public health agencies need to cope with dengue outbreaks efficiently. The implementation of a surveillance system is one strategy that can be used to monitor the epidemics of dengue, which may help to mitigate or reduce the disease outbreaks. Epidemiological surveillance systems gather, analyze, and interpret data about a particular disease and the results will be reported to the relevant public health authorities. However, delayed reporting of dengue cases, due to the time-consuming diagnosis process to confirm cases, may render existing surveillance systems ineffective.

Researchers have studied how viruses spread and predicted the distribution of diseases in different regions by modeling and simulating the disease epidemics. This has helped relevant public health agencies to implement active response plans, improve early detection of the outbreak, controlling mosquito populations, migrating at-risk people, and preparing enough physicians and hospitals to respond to the epidemic (Focks et al. 1995; Derouich & Boutayeb 2006; Medeiros et al. 2011). However, accurate modeling and simulation depends on the data, models, and parameters used, to which simulation results are very sensitive. For instance, a minor change of one particular parameter may result in drastically different results. In addition, some experiments may require high performance computing for modeling and simulation of the disease epidemics; otherwise, experiments can not be easily completed within reasonable time frames (Keeling & Ross 2008; Mikler et al. 2009; Bisset et al. 2010).

Another technique to prepare for the spread of infectious diseases is the development of surveillance systems that utilize data from other resources, including websites or social networks. Twitter, for instance, is a type of online social network with data that has been used to predict disease epidemics using different methods in order to help epidemiologists and public health organizations control the disease outbreak (Gomide et al. 2011; Signorini et al. 2011). The use of social network data can help researchers estimate the number of such dengue incidences ahead of official reports, which are often delayed due to the laboratory process. The results of this type of study can help researchers make reasonable predictions about the number of dengue cases at near real-time even though the people who tweeted did not go to hospitals or get diagnosed. This approach is called syndromic surveillance, which is conducted to

quickly detect and monitor disease epidemics before diagnoses are confirmed. This will help prepare public health workers and medical facilities to accommodate potential patients, thereby reducing morbidity and mortality (Buehler et al. 2008; Henning 2008).

The use of social media to advance the epidemiology of Influenza is a good example of improving disease surveillance. Corley et al. collected data from *Spinn3r* during the Fall flu season from August 1st to October 1st, 2008. After the data was mined and analyzed, the results showed a high correlation between the data on *Spinn3r* and the US Center for Disease Control and Prevention surveillance reports ($r = 0.767$) with 95% confidence (Corley et al. 2009).

Twitter data from social networks are another resource that can be used for monitoring disease epidemics. Researchers from the United Kingdom, for instance, analyzed Twitter data for 24 weeks during the occurrence of the seasonal flu H1N1 in 2009. The results indicated that the Twitter data can be used to monitor flu activity in regions. The correlation between the normalized Twitter data and the UK Health Protection Agency (HPA) for five regions was over 0.8 (Lamos & Cristianini 2010). In addition, methodological approaches for utilizing tweets for disease surveillances have been widely used to monitor not only Influenza H1N1, but also many other diseases including malaria, dengue, yellow fever, measles, poisoning, cholera, typhoid, hepatitis, and smallpox (Achrekar et al. 2011; Gomide et al. 2011; Kriek et al. 2011; Signorini et al. 2011).

Regarding the effort of utilizing the social networking data from Facebook, which has become a popular social network, many studies have conducted experiments to gather Facebook data that people posted in public for analyzing, monitoring, and predicting trends (Chu et al. 2011; Cyjiki & Michahelles 2011; Nguyen & Tran 2011). Fan and Yeung collected Facebook data for modeling virus propagation. The findings of this research show that people posted information about their illnesses in their messages on Facebook for entertainment. This indicates that Facebook data can be used for monitoring disease outbreaks (Fan & Yeung 2010).

Data from social networks in the previous examples indicate that such data can be used in syndromic surveillance systems: the systems can track disease epidemics, predict the number cases, and identify the disease outbreak at near real-time. In addition, Achrekar et al. introduced the framework called Social Network Enabled Flu Trends (SNEFT), which performs a crawling over the Twitter messages that are related to the H1N1 symptoms and the data from the influenza-like illness (ILI) reports (this report is always delayed by 1–2 weeks). Researchers collected 4.7 million tweets from Oct 18th, 2009, to Oct 31st, 2010, and used the auto-regression model to analyze the data and predict the ILI incidences.

The correlation of the two sets of data was very high ($r = 0.98$), indicating that Twitter data are useful data for disease surveillance (Achrekar et al. 2011).

Different data mining or machine learning algorithms, such as Naive Bayes, Decision Tree, Artificial Neural Networks, Support Vector Machine, can be used for analyzing disease data in order to identify the high season of outbreaks and to predict the number of cases (Chakoumakos 2012). In previous work, researchers predicted the occurrence of heart disease by analyzing data in the fields of medicine, computer science, and engineering from journals and publications provided on the Internet. Different algorithms were used; the results showed that Decision Tree outperforms Naives Bayes and Artificial Neural Networks (Soni et al. 2011).

In a second study, researchers collected healthcare data from a database at the University of California at Irvine. The data consists of different attributes related to heart disease, such as age, sex, blood pressure, and blood sugar. These data have been analyzed using multiple algorithms, including Rule Based, Decision Tree, Naive Bayes, and Artificial Neural Networks. The findings showed that the Naive Bayes model outperformed the others. The model could predict a heart attack with the accuracy of 84% (Srinivas et al. 2010).

The efforts in disease surveillances have been extended to dengue fever. For instance, Gomide and others collected data from two different sources in Brazil: the Twitter data that are related to dengue terms (dengue) and the official dengue cases. The linear regression model was used for predicting the number of dengue cases. The results showed that the value of predictive model and the Twitter data have high correlation with $R^2 = 0.9578$ (Gomide et al. 2011).

Our research demonstrates the use of data from another resource on the Internet: typical search query data that people entered in the Google search engine. The data were analyzed to predict the number of dengue cases in Thailand. Several machine learning techniques were applied to find the best type of predictive model. Additionally, the query search terms that related to both dengue search terms and climate search terms, including rainfall, temperature, and humidity terms, were collected to find whether these terms can be used to predict the number of dengue cases.

4.3.2 Methodology for dengue cases prediction

4.3.2.1 Framework

As shown in Fig. 4.4, the design of our study consists of five components, which are search query data, official data, predictive models, models validation, and

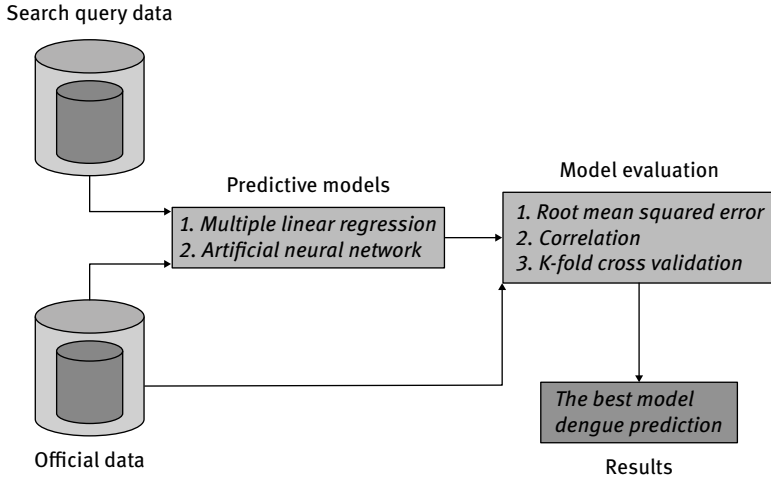


Fig. 4.4: Research framework.

results. The data from the first two components are used to build predictive models that forecast the number of dengue cases. The validation component measures the goodness or the performance of these models using root mean squared error (RMSE), Person correlation coefficient (r), and K-fold cross validation. Finally, the results shall identify the best type of predictive model.

4.3.2.2 Data sets

This study collected two types of data sources, search query data (or search terms) and official data.

(1) Search query data

The frequencies of different search queries were collected from the Google Trends web search engine, which is an online search tool that shows how often a particular search term has been queried over periods of time, different regions, and various languages (Rouse 2012). This study focused on people in Thailand who used the Google search engine to find dengue related information. Therefore, the collection of search terms have been selected from both languages, English and Thai. These terms are related to dengue, climate, and temperature, which have been shown to affect the vector dynamics and thus the severity of dengue epidemic. Search query data were collected from January 2008 to the end of August 2013. The collected data were categorized into four groups as shown in Tab. 4.1.

Tab. 4.1: Variables and search queries.

Category	Variable	Search query	Remark
Dengue terms	X1	dengue	
	X2	dengue fever	
	X3	dengue symptoms	
	X4	mosquito bites	
	X5	ยุงลาย	[Thai] aedes aegypti
	X6	ไข้เลือดออก	[Thai] dengue
	X7	โรคไข้เลือดออก	[Thai] dengue fever
	X8	อาการไข้เลือดออก	[Thai] dengue symptoms
	X9	ยุงกัด	[Thai] mosquito bites
	X10	ป้องกันไข้เลือดออก	[Thai] dengue prevention
Rainfall terms	X11	rain	
	X12	raining	
	X13	rainy season	
	X14	flood	
	X15	น้ำฝน	[Thai] rain
	X16	ฝนตก	[Thai] raining
	X17	ฤดูฝน	[Thai] rainy season1 (formal term)
	X18	หน้าฝน	[Thai] rainy season2 (informal term)
	X19	น้ำท่วม	[Thai] flood
	X20	น้ำขัง	[Thai] waterlogging
	X21	ปริมาณน้ำฝน	[Thai] rainfall amount
Temperature & humidity terms	X22	temperature	
	X23	hot	
	X24	humid	
	X25	humidity	
	X26	อุณหภูมิ	[Thai] temperature
	X27	ร้อน	[Thai] hot
	X28	ชื้น	[Thai] humid
	X29	ความชื้น	[Thai] humidity
Concept terms	X30	[Concept] dengue	X1+X2+X6+X7
	X31	[Concept] dengue symptoms	X3+X8
	X32	[Concept] mosquito bites	X4+X9
	X33	[Concept] rain	X11+X12+X15+X16
	X34	[Concept] rainy season	X13+X17+X18
	X35	[Concept] flood	X14+X19+X20
	X36	[Concept] temperature	X22+X26
	X37	[Concept] hot	X23+X27
	X38	[Concept] humidity	X24+X25+X28+X29

Note that some search terms have a very low frequency. For instance, the frequency of *dengue* in 2008 was zero, which would cause low accuracy of results. Hence, it is necessary to combine those search terms that have the same or similar meaning. These combined terms are referred to as *concept* terms. For instance, the [Concept] – *dengue* is the combination of *dengue*, *dengue fever*, [Thai] *dengue*, and [Thai] *dengue fever*.

(2) Official data

The official data, which count the reported dengue cases in Thailand from January 2008 to the end of August 2013 was collected from Department of Disease Control, Thailand. The information is hosted and freely available from the official website: <http://www.thaivbd.org/dengue.php?id=234>.

4.3.2.3 Predictive models

(1) Multiple linear regression

Multiple linear regression is the statistical method employed to model the relationship between a dependent variable (response variable) and two or more independent variables (explanatory variables) (Mendenhall et al. 1993; Buntinas & Funk 2005; Sullivan & Verhoosel 2010). This technique attempts to measure the strength of the relationship among the variables. A simple equation of the multiple linear regression model is in the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (1)$$

where y is the response variable, β_0 denotes the intercept of the regression line, β_1 to β_k are coefficients or slopes, k is the number of independent variables, ε is the error term or noise, and x_1 to x_k are independent variables or predictors. In this study, each independent variable (search term) represents the volume of the search query and the dependent variable represents the official number of dengue cases.

(2) Artificial neural network

An artificial neural network or neural network is a machine learning model inspired by the biological nervous system, especially the human brain which consists of nerve cells called neurons. A neuron is connected to other neurons via axons, which transmit nerve impulses to other neurons. A neuron connects to axons via dendrites, which receive signals from other neurons (Zhang et al. 1998; Tan et al. 2005).

Figure 4.5 shows a multilayer perceptrons approach (MLP), which consists of several layers of nodes equivalent in a small scale to the many neurons in a human brain. The first layer is an input layer, which receives external information. Each node in this layer represents an independent variable. The last layer is the output layer, which represents the dependent variable. Between the input and output layer there are one or more intermediate layers called hidden layers. Hidden layers contain the nodes that connect the input layer with the output layer by means of weighted links. The different weights in different links represent the strengths of the relationships among the neurons. For instance,

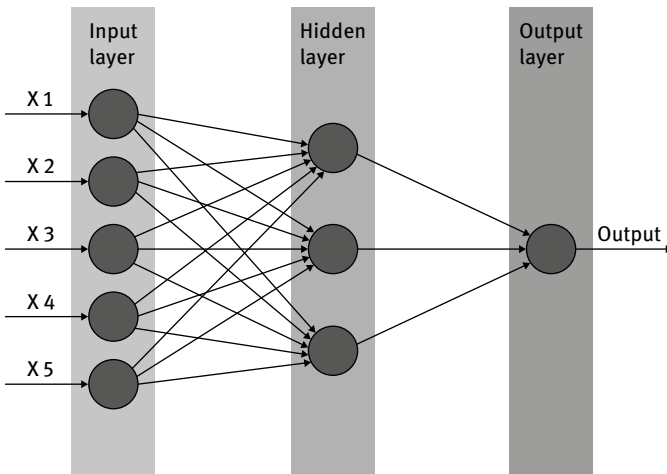


Fig. 4.5: A typical multilayer feed-forward artificial neural network (ANN).

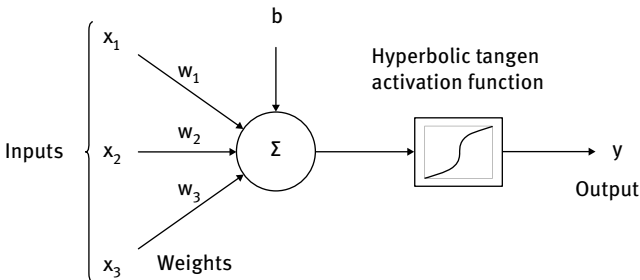


Fig. 4.6: Architecture of applied neural network.

a strong connection between any two neurons is determined by a high weight value of the link between the neurons (Kaastra & Boyd 1996; Zhang et al. 1998).

With respect to the training process to learn the weights, the feed-forward and backpropagation algorithms are the most commonly used (Kaastra & Boyd 1996). The algorithms start with the feed forward phase. Input entry data is fed into the network and the weight values in each link are initialized randomly. The next phase compares the output with a desired value (corresponding target) and feedback error (backpropagation), then updates the new weights whenever necessary.

In the feed forward approach, each node in the hidden and the output layers has its weighted sum (or net sum), which is the summation of the products of the weight and the neuron in the previous layer. The equation of the weighted sum is defined as

$$X_j = \sum_{i=1}^n (w_{ij} X_i) + b \quad (2)$$

where X_j represents the weighted sum of the j th neuron in the hidden and output layers. n denotes those neurons forming the previous layer and X_i is the output of the i th neuron in the previous layer. w_{ij} is the weight between the j th and i th neuron, and b represents the sum function, which computes the effect of inputs and weights (Ali & Tohid 2012).

The next step of the feed forward method is to apply the activation function in order to reduce the weighted sum values into small numbers. It is the activation function that prevents the neural networks from consuming excessive training time due to large values in the outputs from the weighted sums (Kaastra & Boyd 1996; Akintola et al. 2011). Activation functions commonly used to do this are sigmoid (logistic), hyperbolic tangent (tanh), sine or cosine, and linear (identity) functions (Zhang et al. 1998; Tan et al. 2005).

In this study the hyperbolic tangent function is used by the neurons in the hidden layer. The hyperbolic tangent function is defined as

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3)$$

where $\sinh x$ is the hyperbolic sine of x , $\cosh x$ is the hyperbolic cosine of x , and e is Euler's number (approximately 2.718281828).

In addition, the activation function for the output layer used in this study is linear function ($y = f(x)$) because the expected value is the number of dengue cases (continues values).

4.3.2.4 Validation

(1) Root mean squared error (RMSE)

The root mean squared error is one of the most common measurement methods to evaluate the goodness of different models (Mendenball et al. 1993). This method measures the performance of candidate models by taking the standard deviation of the prediction errors (residuals). The lowest value of RMSE gives the best model because the lowest value indicates less residual variance. The RMSE can be computed as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y})^2} \quad (4)$$

where y_i is the actual value (official number of dengue cases), \hat{y} is the prediction value, and n is the size of data.

The RMSE is very sensitive with large errors that would result in a very high value (Mendenball et al. 1993). Normalized RMSE is used to transform the RMSE into a small value. The equation of normalized RMSE is defined as

$$NRMSE = \frac{RMSE}{X_{max} - X_{min}} \quad (5)$$

where $NRMSE$ is the normalized root mean squared error, $RMSE$ is the root mean squared error, and X_{max} and X_{min} are the maximum and the minimum of the actual (observed) values, which come from the official data.

(2) Pearson correlation coefficient (r)

The Pearson coefficient measures the correlation between two variables. It determines how strong the relationship between two variables is. The following equation defines the Pearson coefficient:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}} \quad (6)$$

where x_1, \dots, x_n are the values of independent (predictor) variables, \bar{x} denotes the mean of x_1, \dots, x_n , y_1, \dots, y_n , which are the values of dependent variables, and \bar{y} represents the mean of y_1, \dots, y_n . The values of r are between -1 and 1 , $r = 1$ indicates perfect positive correlation, $r = -1$ indicates perfect negative correlation, and $r = 0$ indicates the absence of correlation between the two random variables (x and y). In this study, x_1, \dots, x_n are predicted values (cases) and y_1, \dots, y_n represent reported cases.

(3) K-fold cross-validation

Cross-validation is used to evaluate and compare the learning algorithms by dividing the corpus into two data sets: training and testing. The training set is used to learn or to train, whereas the testing set is used to validate the model. In the k-fold cross-validation, the data corpus is partitioned into k equally sized subsets. During each execution, one subset is chosen for testing while the others are used for training. The process repeats k times in order to allow each data subset to be used as a testing set (Tan et al. 2005).

In this study, both official dengue cases and search query data were partitioned into six subsets, 2008–2013. Each subset consists of 12 months, except the data from 2013, which was available from January to the end of August. Although this data set did not cover a complete year, it is acceptable as it spans over most of the year.

4.3.3 Prediction analysis

In this study, IBM SPSS was used to analyze the corpus, which was partitioned into six data subsets. Each subset (data for a single year) was used for testing, and the rest of the subsets were used for training. Having four categories per year, the analysis produced 24 models.

4.3.3.1 Multiple linear regression

In order to build the models to predict the number of dengue cases, multiple linear regression was used for analysis. SPSS was executed multiple times based on the different subsets of the corpus and the categories. For analysis, the “backward” method was used in the process.

The “backward” method starts with all candidate variables (search terms) then removes the variable which improves the model.⁴ This process is repeated until no improvement can be made (all the remaining variables have a p -value less than or equal to 0.05). The results of the multiple linear regression analysis are summarized in Tab. 4.2.

Nineteen search terms shown in Tab. 4.2 are the terms used in the fitted models, which were built from multiple linear regression analysis. As can be seen from this table, there are six dengue search terms, seven rainfall terms, two temperature terms, and five concept terms that were used to predict the number of dengue cases.

For dengue search terms, there are two English terms and three Thai terms that were produced by the analysis. In 2008, only two terms (*[Thai] dengue* and *[Thai]*

⁴ The term that yield the least for the fitted model.

dengue prevention) were used in the fitted model, indicating that the number of people who used dengue related search terms in 2008 was not high. In 2009, three dengue terms were used in the fitted model; the correlation for the training set is

Tab. 4.2: Variables, or search terms, produced from multiple linear regression analysis, the correlations (*r*) between the predicted values and the reported dengue cases, and the normalized root mean squared error (*NRMSE*). The best prediction of each model are mostly concept terms models, indicating that the concept terms improve in the predictions.

Category	Variable	Search term	Year					
			2008	2009	2010	2011	2012	2013
Dengue terms	X1	dengue		✓	✓	✓	✓	✓
	X3	dengue symptoms			✓			
	X6	[Thai] dengue	✓	✓	✓	✓	✓	✓
	X7	[Thai] dengue fever		✓	✓	✓	✓	✓
	X10	[Thai] dengue prevention	✓		✓	✓	✓	✓
		<i>r – training</i>	0.92	0.90	0.79	0.90	0.90	0.83
		<i>r – testing</i>	0.85	0.34	0.94	0.88	0.74	0.96
		<i>NRMSE – testing</i>	0.608	0.459**	0.257	0.195***	0.290	0.239
Rainfall terms	X11	rain	✓			✓		
	X13	rainy season	✓		✓	✓	✓	✓
	X16	[Thai] raining	✓	✓	✓	✓	✓	✓
	X17	[Thai] rainy season1	✓	✓	✓	✓	✓	✓
	X18	[Thai] rainy season2		✓	✓	✓		✓
	X19	[Thai] flood			✓			
	X20	[Thai] waterlogging	✓	✓			✓	✓
		<i>r – training</i>	0.86	0.87	0.88	0.88	0.85	0.79
		<i>r – testing</i>	0.83	0.80	0.85	0.64	0.68	0.93
	<i>NRMSE – training</i>	0.237	0.768	0.239	0.747	0.374	0.225*	
Temperature & humidity terms	X26	[Thai] temperature	✓	✓	✓	✓	✓	✓
	X29	[Thai] humidity				✓	✓	✓
		<i>r – training</i>	0.49	0.52	0.52	0.56	0.56	0.56
		<i>r – testing</i>	0.64	0.55	0.93	0.70	0.75	0.87
		<i>NRMSE – testing</i>	0.293*	0.650	0.341	0.454	0.505	0.462

(Continued)

Tab. 4.2: (Continued)

Category	Variable	Search term	Year					
			2008	2009	2010	2011	2012	2013
Concept terms	X30	[Concept] dengue	✓	✓	✓	✓	✓	✓
	X31	[Concept] dengue symptoms	✓	✓	✓	✓	✓	✓
	X34	[Concept] rainy season	✓	✓	✓	✓	✓	✓
	X35	[Concept] flood	✓	✓	✓		✓	✓
	X36	[Concept] temperature			✓			
		<i>r – training</i>	0.90	0.92	0.93	0.91	0.91	0.84
		<i>r – testing</i>	0.94	0.82	0.89	0.56	0.79	0.95
	<i>NRMSE – testing</i>	0.155**	0.555	0.221**	0.305	0.261**	0.141***	

*Represents the best prediction of the model based on categories.

**Represents the best prediction of the model based on years.

***Represents the best prediction of the model based on categories and years.

very strong ($r = 0.90$), but there is a low value for testing ($r = 0.34$). This may be caused by a weak relationship between the data for testing and the number of reported dengue cases. In 2010 and 2012, the correlations for testing are lower than those for training. The large size of the dengue outbreak that occurred in 2010 and 2012 may be a plausible explanation.

For the category of rainfall terms, there are seven search terms (out of eleven terms) that have been used in the fitted models: two English terms and five Thai terms. The correlations around 0.80 indicates that the predicted values and the official dengue cases have a strong relationship.

This suggests that the rainfall terms can be used to predict number of dengue cases. However, the moderate correlations in 2011 and 2012 ($r = 0.64$ and 0.68 , respectively) are lower than those obtained for training. These may be due to an abundance of rainfall in 2011, which caused significant flooding which lasted until the beginning of 2012. Due to this anomaly, an overestimation is expected in the predictions in these 2 years.

With eight terms of temperature and humidity, there are only two terms ([*Thai*] *temperature* and [*Thai*] *humidity*) used in the fitted models. In fact, there is only one term ([*Thai*] *temperature*) used in the predictive models in 2008–2010. In addition, the correlations between 0.49 and 0.56 for training data sets indicate moderate relationships between the reported cases and these search terms.

The concept terms, the combination of search terms that have very close meanings, were created to improve the predictive models because of the low frequency of most search terms, which would cause lower correlations for testing when comparing to those correlations for training. As can be seen in Tab. 4.2, the correlations between the predicted values and the observed values are very strong and higher than those in other categories, indicating that the search terms used in this study can be used to predict the number of dengue cases. However, there is a medium correlation ($r = 0.56$) for testing in 2011. This might be due to the high frequency of the term [Concept] *rainy season*, due to the flood in 2011, and therefore, affects the fitted model (overestimation).

In 2013, the correlations are very strong, and the correlation values for testing are higher than the correlation values for training. In 2013 the government and the public health agencies have frequently informed the people about the number of dengue cases since the beginning of the year, leading to the high frequency of all search terms that correspond to the number of dengue cases.

The normalized root mean squared error (NRMSE) for each model was computed to find the best predictive model based on categories and years. The models for concept terms are generally the best predictors for dengue cases, indicating that the concept terms improve the forecasts.

Examples of the results from multiple linear regression for 2012 and 2013 are shown in Fig. 4.7. The predicted values, are plotted against the observed values based on four categories. The white backgrounds in these graphs represent the data for testing (predicted data).

4.3.3.2 Artificial neural network

For this analysis, IBM SPSS was used to build different models to predict the number of dengue cases. For each learning model, we used the same predictors (independent variables) as those used for multiple linear regression. As a result, SPSS produced different predictive models. In Tab. 4.3, the correlations between the reported dengue cases and the predicted data produced from artificial neural network analysis are shown.

As shown in Tab. 4.3, most of the correlations are higher than the correlations from multiple linear regression. For example, the correlations for training for all years and all categories, except *temperature and humidity* terms, range between 0.69 and 0.95, while the correlations for testing range from 0.48 to 0.97. Moreover, the correlations for the *temperature and humidity* terms range from 0.53 to 0.94. These indicate that the predicted values produced from ANN analysis have stronger relationship with the number of dengue cases than the predicted data produced by the multiple linear regression.

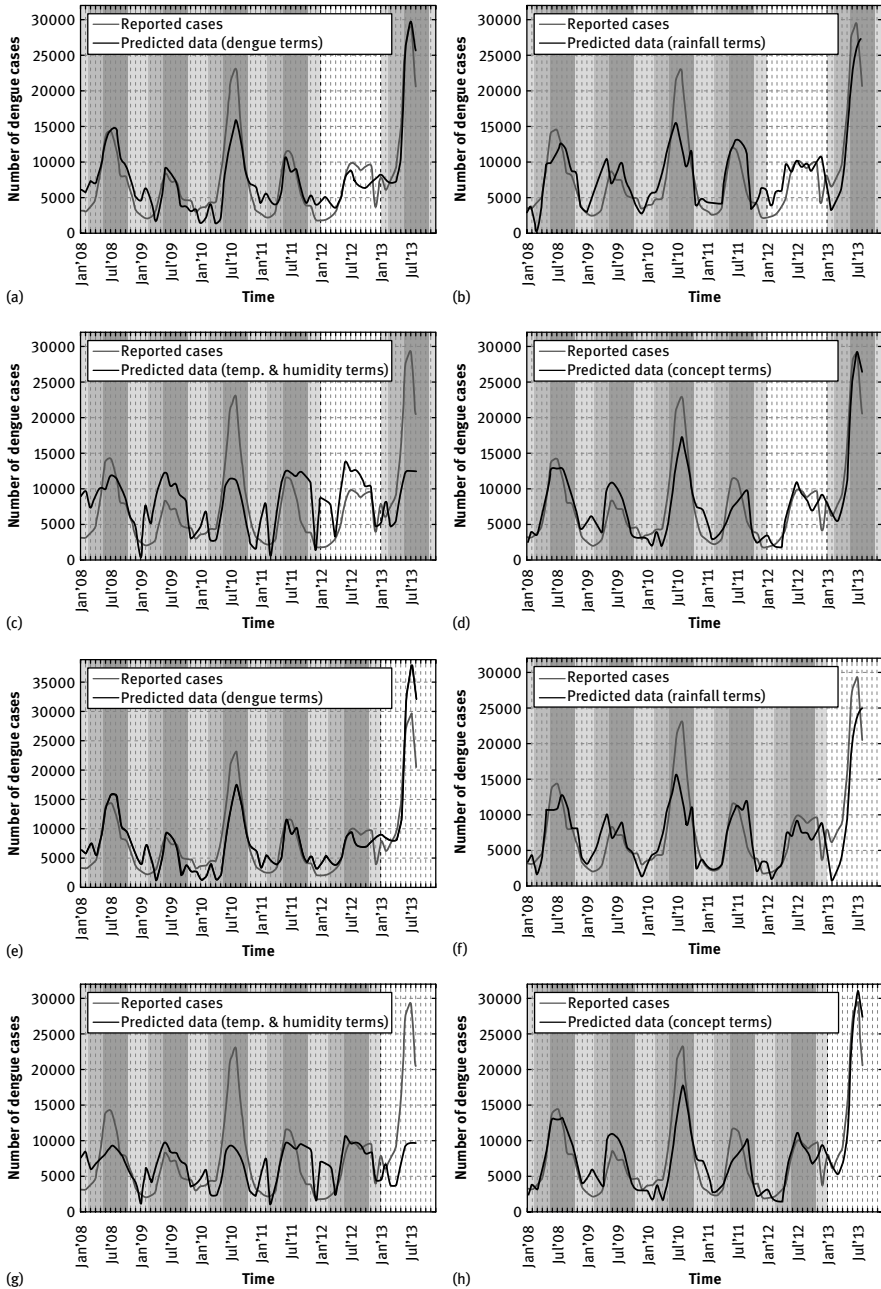


Fig. 4.7: Prediction of dengue cases for 2012 and 2013, white regions, using multiple linear regression mode. Winter (yellow), summer (orange), and rainy (blue) seasons are rendered in the background.

Tab. 4.3: Summary of the artificial neural network analysis. the lowest normalized root mean squared error (NRMSE) for each year and each category indicates the best prediction of the models.

Year	Category	Input layer: # of neurons	Hidden layer: # of neurons	Correlations (<i>r</i>)		
				Training	Testing	<i>NRMSE</i>
2008	Dengue terms	2	1	0.89	0.82	0.290
	Rainfall terms	5	3	0.87	0.84	0.247
	Temperature & humidity terms	1	1	0.53	0.66	0.268*
	Concept terms	4	3	0.92	0.97	0.135***
2009	Dengue terms	3	2	0.93	0.48	0.409
	Rainfall terms	4	3	0.69	0.85	0.456
	Temperature & humidity terms	1	1	0.53	0.59	0.483
	Concept terms	4	3	0.90	0.84	0.274**
2010	Dengue terms	5	3	0.95	0.90	0.168**
	Rainfall terms	5	3	0.93	0.89	0.195
	Temperature & humidity terms	1	1	0.56	0.91	0.326
	Concept terms	5	3	0.95	0.96	0.170
2011	Dengue terms	4	3	0.93	0.89	0.163**
	Rainfall terms	5	3	0.87	0.85	0.300
	Temperature & humidity terms	2	2	0.58	0.76	0.383
	Concept terms	3	2	0.93	0.55	0.294
2012	Dengue terms	4	3	0.94	0.71	0.296
	Rainfall terms	4	3	0.89	0.61	0.329
	Temperature & humidity terms	2	2	0.58	0.78	0.374
	Concept terms	4	3	0.92	0.78	0.259**
2013	Dengue terms	4	3	0.90	0.96	0.106***
	Rainfall terms	5	3	0.84	0.97	0.138
	Temperature & humidity terms	2	2	0.63	0.94	0.424
	Concept terms	4	3	0.81	0.94	0.159

*Represents the best prediction of the model based on categories.

**Represents the best prediction of the model based on years.

***Represents the best prediction of the model based on categories and years.

Examples of the results from the artificial neural network analysis in 2012 and 2013 are shown in Fig. 4.8. The predicted data are plotted against the official dengue cases in four categories. The lowest *NRMSE* values of the models for

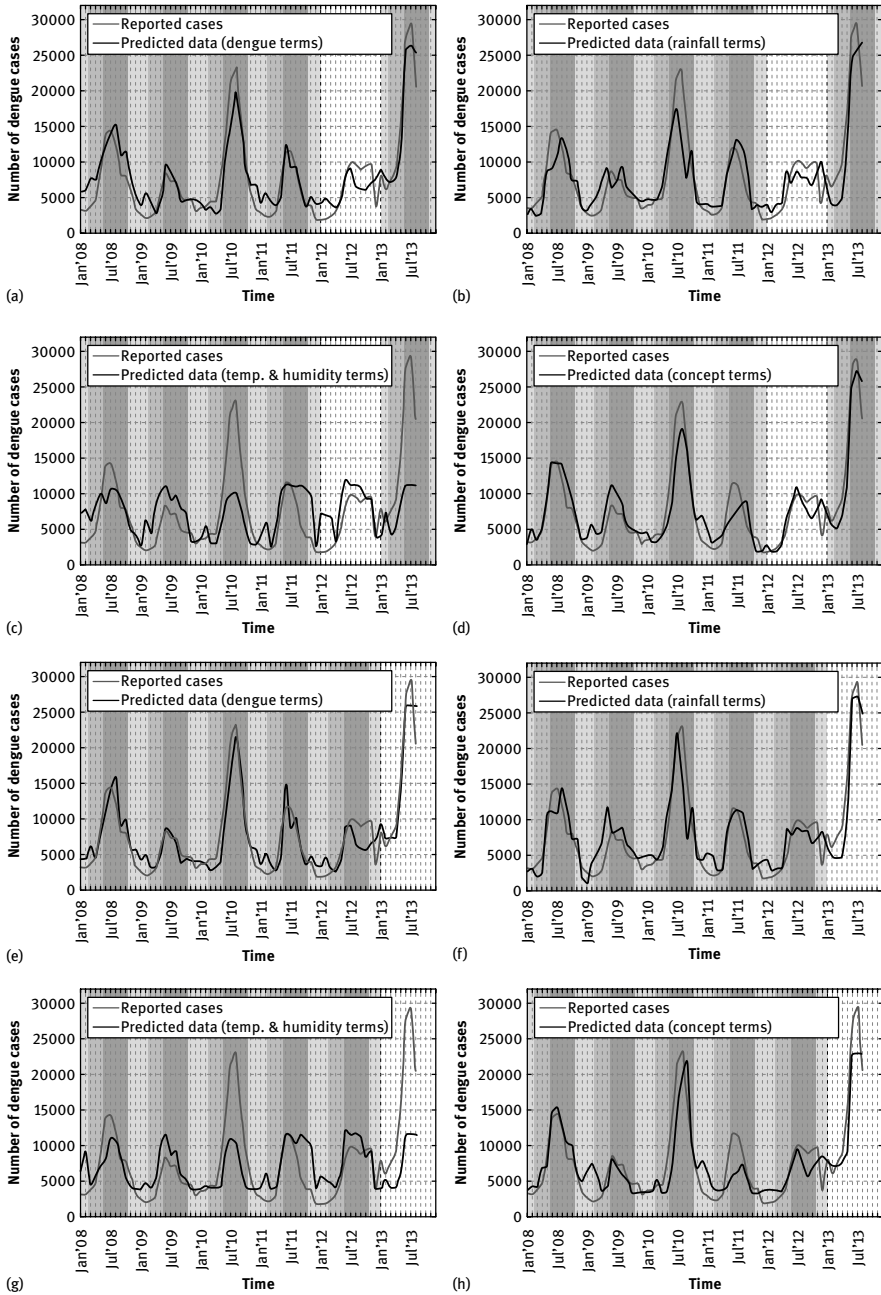


Fig. 4.8: Prediction of dengue cases for 2012 and 2013, white regions, using artificial neural network model. Winter (yellow), summer (orange), and rainy (blue) seasons are rendered in the background.

dengue terms and the models for concept terms indicate the best predictions based on categories and years.

4.3.3.3 Comparison of predictive models

In order to compare the performance of the models created by multiple linear regression and artificial neural network, we assessed the quality of these models by computing the overall correlations (r) and overall *NRMSE*, which evaluates how close the predicted and the observed data are. The highest correlation value represents the strongest relationship between the predicted and observed data, and the lowest value of *NRMSE* indicates the performance of the best predictive model.

Table 4.4 shows the best model for multiple linear regression is the model for concept terms in 2012 with a highest $r = 0.90$ and a lowest total *NRMSE* = 0.0932. The best artificial neural network model is the model in 2013 for dengue terms with a maximum $r = 0.94$ and a minimum total *NRMSE* = 0.0743. On the other hand, the worst multiple linear regression model is the model of temperature and humidity terms in 2010 (lowest $r = 0.50$ and highest total *NRMSE* = 0.1889). In addition, the worst artificial neural network is the model of temperature and humidity terms in 2009 (lowest $r = 0.50$ and highest total *NRMSE* = 0.1877).

Overall, the correlations from artificial neural network (ANN) are mostly higher than those from multiple linear regression (MLR) models, and the total *NRMSE* from artificial neural network (2.8703) is lower than the total *NRMSE* from multiple linear regression. These indicate that the ANN models outperform the multiple linear regression models.

4.3.4 Discussion

The overall correlations of the predicted data and the observed data are mostly higher than 0.80, indicating that these trends are strongly related. It implies that the search terms from Google trends can be used to predict the number of dengue cases. Both the MLR and ANN demonstrated to be good predicting models for the number of dengue cases.

The low frequency of the search terms is directly affecting the fitting and the performance of the models as can be seen in the low correlations of the testing as compared with those of the training. When combining the frequency of search terms that have the same meaning (concept terms), the correlations for testing are mostly higher than those for training. A high correlations in 2013 indicates a strong relationship between the predicted and observed data. This implies that the fitted models work very well and the data for testing are suitable for the prediction. This strong correlation may be due to the fact that Thai public health agencies have issued warnings and informed the population via the media about the ongoing

Tab. 4.4: Comparison of predictive models.

Year	Category	MLR		ANN	
		<i>r</i>	<i>NRMSE</i>	<i>r</i>	<i>NRMSE</i>
2008	Dengue terms	0.83	0.1327	0.77	0.1555
	Rainfall terms	0.85	0.1128	0.86	0.1110
	Temperature & humidity terms	0.50	0.1867	0.54	0.1813
	Concept terms	0.90	0.0942	0.92	0.0854
2009	Dengue terms	0.88	0.1006	0.92	0.0861
	Rainfall terms	0.82	0.1281	0.68	0.1582
	Temperature & humidity terms	0.50	0.1880	0.50	0.1877
	Concept terms	0.89	0.0989	0.89	0.1062
2010	Dengue terms	0.89	0.1011	0.94	0.0768
	Rainfall terms	0.85	0.1144	0.91	0.0914
	Temperature & humidity terms	0.50	0.1889	0.53	0.1833
	Concept terms	0.90	0.0944	0.93	0.0795
2011	Dengue terms	0.90	0.0957	0.93	0.0814
	Rainfall terms	0.77	0.1476	0.84	0.1224
	Temperature & humidity terms	0.54	0.1826	0.56	0.1800
	Concept terms	0.89	0.0982	0.91	0.0887
2012	Dengue terms	0.90	0.0956	0.92	0.0839
	Rainfall terms	0.84	0.1177	0.87	0.1047
	Temperature & humidity terms	0.54	0.1823	0.56	0.1786
	Concept terms	0.90	0.0932	0.92	0.0852
2013	Dengue terms	0.89	0.1111	0.94	0.0743
	Rainfall terms	0.85	0.1161	0.90	0.0939
	Temperature & humidity terms	0.54	0.1865	0.62	0.1732
	Concept terms	0.90	0.0940	0.88	0.1016
Total <i>NRMSE</i>			3.0614		2.8703

dengue outbreak. In addition, the usage of the Thai Internet has been high during this period, consequently leading to a high frequency of search terms in 2013.

The high correlations of rainfall terms indicate a strong relationship of the predicted and observed values. This suggests that rainfall related terms have a significant predictive power and can be utilized to forecast the severity of a dengue epidemic. Nevertheless, the correlations of temperature and humidity terms in all models range between 0.50 and 0.62 indicating that the predicted and the observed data have only a moderate relationship to each other, leading to a decreased accuracy in the prediction of dengue cases. The low values of *NRMSE* stem mostly from the models that used dengue and concept terms. These terms are highly correlated with the number of dengue cases, and they have therefore strong predictive power.

4.4 Conclusion

Our research has focused on a set of specific search terms that have been motivated by the effects of climate change and temperature variation on the mosquito life cycle and consequently the epidemiology of dengue.

Occurrences of climate specific terms together with dengue specific terms in search queries have been obtained from Google trends. Different machine learning techniques have been utilized to derive predictive models that are capable of forecasting the severity of dengue epidemics. The results of our research indicates that Google trend data is in fact suitable for the construction of predictive models which can accurately estimate the incidence of dengue during an epidemic in Thailand.

Some of our research results have been posted on Facebook in an effort to share them with public health professionals in Thailand. In response, some comments have been posted, which express great interest in the model results as shown in Fig. 4.9. It appears that many public health practitioners are seeking support in online social media. We believe that the use of social network sites and the analysis of their contents can drastically improve the prediction of disease incidence in general and dengue cases in particular.

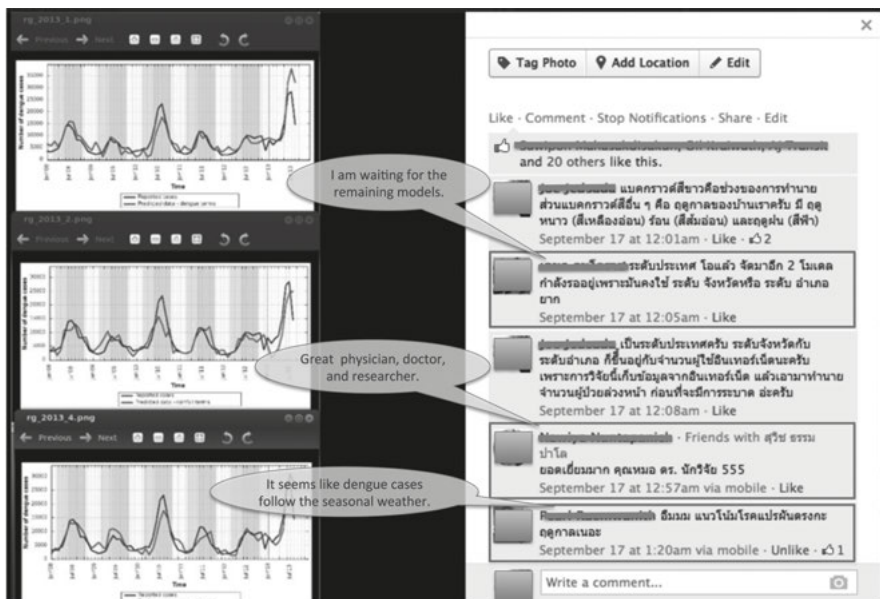


Fig. 4.9: Experiment results posted on Facebook unchained a reaction of positive comments from public health practitioners.

References

- Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. & Liu, B. (2011) 'Predicting flu trends using Twitter data', *IEEE International Workshop on Cyber-Physical Networking Systems*, 702–707.
- Akintola, K. G., Alese, B. H. & Thomson, A. F. (2011) 'Time series forecasting with neural network: a case study of stock prices of intercontinental bank Nigeria'. *IJRRAS*, 467–472.
- Ali, N. & Tohid, A. (2012) 'Prediction the effects of ZnO₂ nanoparticles on splitting tensile strength and water absorption of high strength concrete'. *Mater Res*, 15:440–454. ISSN 1516–1439.
- Altizer, S., Dobson, A., Hosseini, P., Hudson, P., Pascual, M. & Rohani, P. (2006) 'Seasonality and the dynamics of infectious diseases'. *Ecol Lett*, 9(4):467–484.
- Anderson, R. M. & May, R. M. (2002) *Infectious Disease of Humans: Dynamics and Control*. Oxford: Oxford Science Publication.
- Bennet, A. & Joshi, N. (2008) 'Distribution and seasonality of vertically transmitted dengue virus in *Aedes* mosquitoes in arid and semi-arid areas in Rajasthan, India', *J Vector Borne Dis*, 56–59.
- Beserra, E. B., de Castro Júnior, F. P., dos Santos, J. W., da S Santos, T. & Fernandes, C. R. M. (2006) 'Biologia e Exigências Térmicas de *Aedes Aegypti* (L.) (Diptera:Culicidae) Provenientes de Quatro Regies Bioclimáticas da Paraíba', *Neotrop Entom*, 35(6):853–860.
- Bisset, K. R., Chen, J., Feng, X., Ma, Y. & Marathe, M. V. (2010) 'Indemics: an interactive data intensive framework for high performance epidemic simulation', *ICS 2010*, 5:2010.
- Buehler, J. W., Sonricker, A., Paladini, M., Soper, P. & Mostashari, F. (2008) 'Syndromic surveillance practice in the United States: findings from a survey of state, territorial, and selected local health departments', *Adv Dis Surv*, 6(3):1–20.
- Buntinas, M. & Funk, G. M. (2005) *Statistics for the Sciences*. London: Thomson Brooks/cole.
- Carneiro, H. A. & Mylonakis, E. (2009) 'Google trends: a web-based tool for real-time surveillance of disease outbreaks', *Clin Infect Dis*, 49(10):1557–1564.
- CDC. (2010) Dengue – Epidemiology. [Online]. Available: <http://www.cdc.gov/dengue/epidemiology/index.html>.
- CDC. (2013) Division of vector-borne diseases. [Online]. Available: <http://www.cdc.gov/ncezid/dvbd/>.
- Chakoumakos, R. (2012) *Predicting Outbreak Severity Through Machine Learning on Disease Outbreak Reports*, Palo Alto, CA: Stanford University Press.
- Choi, H. & Varian, H. (2012) 'Predicting the present with Google trends', *Econ Trend*, 88:2–9. Jun.
- Chu, H., Deng, D. & Park, J. H. (2011) 'Live data mining concerning social networking forensics based on a Facebook session through aggregation of social media', *IEEE J Select Area Commun*, 29(7):1368–1376.
- Corley, C. D., Miller, A. R., Singh, K. P. & Cook, D. J. (2009) 'Monitoring influenza trends through mining social media', *International Conference on Bioinformatics & Computational Biology*, Las Vages, NA.
- Cyjijiki, I. P. & Michahelles, F. (2011) Intelligent heart disease prediction system using data mining techniques. *IEEE 9th International Conference on Dependable, Autonomic and Secure Computing*.
- Derouich, M. & Boutayeb, A. (2006) 'Dengue fever: mathematical modeling and computer simulation', *Appl Maths Commun*, 177(2):528–544.
- Fan, W. & Yeung, K. H. (2010) Virus propagation modeling in Facebook. *IEEE International Conference on Advances in Social Networks Analysis and Mining*.

- Farnesi, L. C., Martins, A. J., Valle, D. & Rezende, G. L. (2009) 'Embryonic development of *Aedes Aegypti* (Diptera: Culicidae): influence of different constant temperatures', *Mem Inst Oswaldo Cruz*, 104(1):124–126.
- Focks, D. A., Daniels, E., Haile, D. G. & Keesling, J. E. (1995) 'A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results', *Am Soc Trop Med Hyg*, 53(5):489–506.
- Gloor, P. A., Krauss, J., Nann, S., Fischbach, K. & Schoder, D. (2009) Web Science 2.0: Identifying trends through semantic social network analysis. In *Computational Science and Engineering, 2009. CSE'09, International Conference on*, pp. 215–222.
- Gomide, J., Veloso, A., Maria, W., et al., editors. (2011) *Dengue surveillance based on a computational model of spatio-temporal locality of Twitter*, June 14–17. *Proceedings of the ACM WebSci'11*. Koblenz, Germany, ACM.
- Guha-Sapir, D. & Schimmer, B. (2005) 'Dengue fever: new paradigms for a changing epidemiology', *BioMed Central*, 2(1).
- Health Map. (2012) Dengue map. [Online]. Available: <http://www.healthmap.org/dengue/index.php>.
- healthmap.org. (2013) Health Map – Dengue. [Online]. Available: <http://www.healthmap.org/dengue/>.
- Henning, K. J. (2008) Overview of syndromic surveillance what is syndromic surveillance? [Online]. Available: www.cdc.gov/MMWR/preview/mmwrhtml/su5301a3.htm.
- Jansen, B. J., Spink, A. & Saracevic, T. (2000) 'Real life, real users, and real needs: a study and analysis of user queries on the web', *Inform Process Manag*, 36(2):207–227. Mar.
- Kaastra, I. & Boyd, M. (1996) 'Designing a neural network for forecasting financial and economic time series'. *Neurocomputing*, 10:215–236.
- Keeling, M. J. & Ross, J. V. (2008) 'On methods for studying stochastic disease dynamics', *J R Soc Interface*, 5(19):171.
- Krause, R. M. (1997) *Emerging Infections*. Academic Press, 15 East 26th St., 15th Floor, New York, New York 10010, USA, first edition.
- Kriek, M., Dreesman, J., Otrusina, L. & Denecke, K. editors. (2011) *A new age of public health: identifying disease outbreaks by analyzing Tweets*. Health WebScience Workshop, Proceedings of the ACM WebSci'11. Koblenz, Germany, ACM.
- Lamos, V. & Cristianini, N. (2010) Tracking the flu pandemic by monitoring the Social Web. *IEEE International Conference Workshop on Conitive Information Processing*.
- Madoff, L. C., Fisman, D. N. & Kass-Hout, T. (2011) 'A new approach to monitoring dengue activity', *PLoS Negl Trop Dis*, 5(5):e1215.
- Medeiros, LCdC., Castilho, C., Braga, C., et al. (2011) 'Modeling the dynamic transmission of dengue fever: investigating disease persistence', *PLoS Negl Trop Dis*, 5(1):e942.
- Mendenhall, W., Reinmuth, J. & Beaver, R. (1993) *Statistics for Management and Economics*. Duxbury press.
- Mikler, A. R., Bravo-Salgado, A. & Corley, C. D. (2009) 'Global stochastic contact modeling of infectious diseases', *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, (5), doi: 10.1109/IJCBS.2009.84.
- Nguyen, K. & Tran, D. A. (2011) An analysis of activities in Facebook. In *IEEE 8th consumer communications and networking conference-emerging and innovative consumer technologies*.

- Palaniappan, S. & Awang, R. (2008) 'Intelligent heart disease prediction system using data mining techniques', In *Computer Systems and Applications*. IEEE/ACS International Conference. March 31–April 4.
- Potts, J. A., Gibbons, R. V., Rothman, A. L., et al. (2010) 'Prediction of dengue disease severity among pediatric Thai patients using early clinical laboratory indicators', *PLoS Negl Trop Dis*, 4(8):e769.
- Racloz, V., Ramsey, R., Tong, S. & Hu, W. (2012) 'Surveillance of dengue fever virus: a review of epidemiological models and early warning systems', *PLoS Negl Trop Dis*, 6(5):1–9.
- Rouse, M. (2012) Google trends. [Online]. Available: <http://whatis.techtarget.com/definition/Google-Trends>.
- Shankman, S. (2012) The top travel trends of 2012 as defined by Google searches. [Online]. Available: <http://skift.com/2012/12/12/travel-trends-in-2012-as-defined-by-google-searches/>. Dec.
- Sharp, T. M., Hunsperger, E., Santiago, G. A., Muñoz-Jordan, J. L., Santiago, L. M., Rivera, A., Rodríguez-Acosta, R. L., Feliciano, L. G., Margolis, H. S. & Tomashek, K. M. (2013) 'Virus-Specific differences in rates of disease during the 2010 dengue epidemic in Puerto Rico', *PLoS Negl Trop Dis* Apr.
- Shuzhen, S., Ramirez, J. L., & Dimopoulos, G. (2012) 'Dengue virus infection of the *Aedes aegypti* salivary gland and chemosensory apparatus induces genes that modulate infection and blood-feeding behavior', *PLoS Pathog*, 8(3):e1002631, 03 2012. doi: 10.1371/journal.ppat.1002631.
- Signorini, A., Segre, A. M. & Polgreen, P. M. (2011) 'The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic', *PLoS One*, 6(5):1–9.
- Soni, J., Ansari, U. & Sharma, D. (2011) 'Predictive data mining for medical diagnosis: an overview of heart disease prediction', *Int J Compt Appl*, 17(8):43–48.
- Srinivas, K., Rani, B. K. & Govrdhan, A. (2010) 'Applications of data mining techniques in healthcare and prediction of heart attacks', *Int J Comput Sci Eng*, 2(2):250–255.
- Sugiyama, K., Hatano, K. & Yoshikawa, M. (2004) 'Adaptive web search based on user profile constructed without any effort from users', *WWW' 04 Proceedings of the 13th International Conference on World Wide Web*.
- Sullivan, M & Verhoosel, J. C. M. (2010) *Statistics: Informed Decisions Using Data*. Upper Saddle River, NJ: Prentice-Hall.
- Tan, P., Steinbach, M. & Kumar, V. *Introduction to data mining*. Addison-Wesley, 2005.
- The Dengue Vaccine Initiative. (2014) About DVI. [Online]. Available: <http://www.denguevaccines.org/>. Jan.
- Watts, D. M., Burke, D. S., Harrison, B. A., Whitmire, R. E. & Nisalak, A. (1987) 'Effect of temperature on the vector efficiency of *Aedes aegypti* for dengue 2 virus'. *Am J Trop Med Hyg*, 36(1):143–152.
- WHO. (2010) Dengue in the Western Pacific Region. [Online]. Available: http://www.wpro.who.int/health/_topics/dengue/.
- WHO. (2012) Dengue and severe dengue. [Online]. Available: <http://www.who.int/mediacentre/factsheets/fs117/en/>.
- Xiao-Xia, G. U. O., Tong-Yan, Z., Yan-De, D., Shu-Nan, J. & Bao-Lin, L. U. (2004) 'Transmission of dengue 2 virus by diapausing eggs of *Aedes albopictus*', *Acta Entom Sinica*, 47(4):424–428, ISSN 04546296. URL <http://www.insect.org.cn/EN/column/column105.shtml>.
- Zhang, G., Patuwo, B. E. & Hu, M. Y. (1998) 'Forecasting with artificial neural network: the state of the art', *Int J Forecast*, 14:35–62.

Kambiz Ghazinour, Marina Sokolova and Stan Matwin

5 A study of personal health information posted online: using machine learning to validate the importance of the terms detected by MedDRA and SNOMED in revealing health information in social media

Abstract: With the increasing amount of personal information that is shared on social networks, it is possible that the users might inadvertently reveal some personal health information. In this work, we show that personal health information can be detected and, if necessary, protected. We present empirical support for this hypothesis, and furthermore we show how two existing well-known electronic medical resources MedDRA and SNOMED help to detect personal health information (PHI) in messages retrieved from a social network site, MySpace. We introduce a new measure – risk factor of personal information – that assesses the likelihood that a term would reveal personal health information. We synthesize a profile of a potential PHI leak in a social network, and we demonstrate that this task benefits from the emphasis on the MedDRA and SNOMED terms. Our study findings are robust in detecting sentences and phrases that contain users' personal health information.

5.1 Introduction

Studies of personal health information (PHI) posted on public communication hubs (e.g., blogs, forums, and online social networks) rely on four technologies: privacy preserving data mining; information leakage prevention; risk assessment; and social network analysis.

PHI relates to the physical or mental health of the individual, including information that consists of the health history of the individual's family, and information about the healthcare provider (Ghazinour, Sokolova & Matwin 2013). We differentiate terms revealing PHI from medical terms that convey health information which is not necessarily personal (e.g., "smoking can cause lung cancer"),

and terms that despite their appearance to be health-related have no medical meaning (e.g., “I have pain in my chest” vs. “I feel your pain”).

We believe that the online social networks’ growth and the general public involvement make social networks an excellent candidate for health information privacy research.

In this chapter we show empirically how personal health information is disclosed in social networks. Furthermore, we show how two existing electronic linguistic medical resources (i.e., MedDRA and SNOMED) help detect personal health information in messages retrieved from a social network.¹ Both resources are well-established medical dictionaries used in biomedical text mining. We use machine learning to validate the importance of the terms detected by these two medical dictionaries in revealing health information and analyze the results of MedDRA (Medical Dictionary for Regulatory Activities) and SNOMED (Systematized Nomenclature of Medicine). We also evaluate our algorithm’s performance by manually finding sentences with PHI that were not detected by our algorithm. Our algorithm gives strong results in terms of the sentences that it detects for containing PHI and only an estimated false negative of 0.003 which is considered a great result for missing PHI compared to the state of the art of the false negative rate of 2.9–3.9 (Miles, Rodrigues & Sevdalis 2013).

In Section 2, we provide background material on the evolution and use of social networks and briefly discuss related work in the area of personal health information and this new form of online communication. Section 3 describes a brief introduction of the main technologies used for this type of research. Section 4 introduces current computational linguistic resources used in medical research. Section 5 explains our empirical study and Section 6 discusses our findings and introduces the *Risk Factor of Personal Information* and contributions of this study. Section 7 discusses how we use machine learning to validate our hypotheses. Section 8 concludes the paper and gives future research directions. Preliminary results of this work were published in (Ghazinour, Sokolova & Matwin 2013).

5.2 Related background

5.2.1 Personal health information in social networks

Social networks can be used for personal and/or professional purposes. Nowadays many healthcare providers are using social networks in their practices to interact with other colleagues, physicians and patients to exchange medical information,

¹ www.eecs.uottawa.ca/~stan/PHI2013data.txt

or to share their expertise and experiences with a broad audience (Keckley & Hoffman 2010). In addition to communicating with healthcare providers, the emergence of social networks, weblogs and other online technologies, has given people more opportunities to share their personal information (Gross & Acquisti 2005). Such sharing might include disclosing personal identifiable information (PII) (e.g., names, address, dates) and personal health information (PHI) (e.g., symptoms, treatments, medical care) among other factors of personal life.

Cyber-dangers in the healthcare sector generally fall into three categories: the exposure of private or sensitive data, manipulation of data, and loss of system integrity.

Websites such as *Patientslikeme* and *Webmd*, Facebook pages such as *Managing Diabetes* (see Fig. 5.1) and many other publically available pages give examples of where adversaries, or person's whose access to one's personal health information would compromise one's financial and social status, can freely access these information.

It has been shown that 19%–28% of all Internet users participate in medical online forums, health-focused groups and communities and visit health-dedicated web sites (Renahy 2008; Balicco & Paganelli 2011). Recently Pew Research Center (Fox 2011) published their findings based on a national telephone survey conducted in August and September 2010 among 3001 adults in the United States. The complete methodology and results are appended to the Pew Research report.

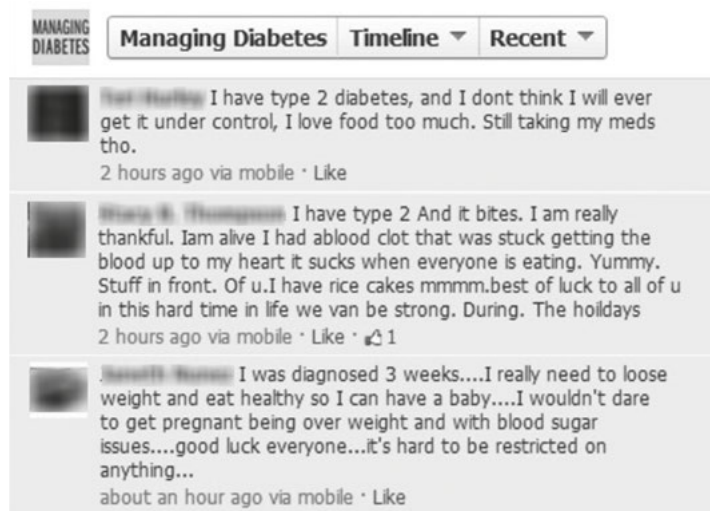


Fig. 5.1: Example of comments on a Facebook page which reveals personal health info.

The survey finds that of the 74% of adults who use the Internet:

- 80% of internet users have looked online for information about any of 15 health topics such as a specific disease or treatment. This translates to 59% of all adults.
- 34% of internet users, or 25% of adults, have read someone else’s commentary or experience about health or medical issues on an online news group, website, or blog.
- 25% of internet users, or 19% of adults, have watched an online video about health or medical issues.
- 24% of internet users, or 18% of adults, have consulted online reviews of particular drugs or medical treatments.
- 18% of internet users, or 13% of adults, have gone online to find others who might have health concerns similar to theirs.
- 16% of internet users, or 12% of adults, have consulted online rankings or reviews of doctors or other providers.
- 15% of internet users, or 11% of adults, have consulted online rankings or reviews of hospitals or other medical facilities.
- Of those who use social network sites (62% of adult internet users, or 46% of all adults):
 - 23% of social network site users, or 11% of adults, have followed their friends’ personal health experiences or updates on the site.
 - 17% of social network site users, or 8% of adults, have used social networking sites to remember or memorialize other people who suffered from a certain health condition.
 - 15% of social network site users, or 7% of adults, have gotten any health information on the sites.

A recent study (Li et al. 2011) had demonstrated a real-world example of cross-site information aggregation that resulted in disclosing PHI. A target patient has profiles on two online medical social networking sites. By comparing the attributes from both profiles, the adversary can link the two with high confidence. Furthermore, the attacker can use the attribute values to get more profiles of the target through searching the Web and other online public data sets. Medical information including lab test results was identified by aggregating and associating five profiles gathered by an attacker, including the patient’s full name, date of birth, spouse’s name, home address, home phone number, cell phone number, two email addresses, and occupation. In fact, using machine learning techniques and algorithms, the extracted health information could be aggregated with other personal information and do more harm than originally expected.

5.2.2 Protection of personal health information

When people share comments or some content related to their personal health information they may share very sensitive information which if combined with their personal identifiable information, can be a dangerous tool in wrong hands. For example, Knudsen (2013) states that 54% of data breaches in the health-care sector were the result of theft. The numbers indicate that providers may be getting better at reducing inadvertent data loss, but criminals have continued to gain an advantage in forcing their way into the world of online communication. From a broader perspective, Harries and Yellowlees (2013) argue that in addition to bomb threat-type attacks, assaults on the public infrastructure – most notably water and power supplies – have the potential to cripple the healthcare system. Another likely target is that bastion of data: the electronic health record.

The success of healthcare providers that use social media and web-based systems is contingent on personal health information provided by individuals (see Fig. 5.2). Focused on the role of personal dispositions in disclosing health information on line, Bansal, Zahedi, and Gefen (2010) demonstrate that individuals' intentions to disclose such information depends on trust and privacy concern, which are determined by personal disposition. Some of the factors that affect personal disposition are personality traits, one's attitude toward



Fig. 5.2: Social media and more ways to share personal health information.

information sensitivity, health status, prior privacy invasions, risk beliefs, and experience which act as intrinsic antecedents of trust. At the same time, uncontrolled access to health information could lead to privacy compromise, breaches of trust, and eventually harm the individuals. As discussed earlier, machine learning and data mining techniques could enable the adversary to gain more information and to do more harm than expected.

Zhang et al. (2011) propose a role prediction model to protect the electronic medical records (EMR) and privacy of the patients. As another example, Miller and Tucker (2009) studied the privacy protection state laws and technology limitations with respect to the electronic medical records.

However, protection of personal health information found in social network postings did not receive as much attention. In part, this is due to the lack of resources appropriate for detection and analysis of PHI in informally written messages posted by the users (Sokolova & Schramm 2011). The currently available resources and tools were designed to analyze PHI in more structured and contrived text of electronic health records (Yeniterzi 2010).

Bobicev et al. (2012) presented results of sentiment analysis in twitter messages that disclose personal health information. In these messages (tweets), users discuss ailment, treatment, medications, etc. They use the author-centric annotation model to label tweets as positive sentiments, negative sentiments or neutral. In another study on twitter data, Sokolova et al. (2013) introduced two semantic-based methods for mining personal health information in twitter. One method uses WordNet (Princeton University 2010) as a source of health-related knowledge, another, an ontology of personal relations. The authors compared their performance with a lexicon-based method that uses an ontology of health-related terms.

5.2.3 Previous work

Some studies analyzed personal health information disseminated in blogs written by healthcare professionals/doctors (Lagu et al. 2008). However, these studies did not analyze large volumes of texts. Thus, the published results may not have sufficient generalization power, (Silverman 2008; Kennedy 2012). Malik and Coulson (2010) manually analyze 3500 messages posted on seven sub-boards of a UK peer-moderated online infertility support group. The results of this study show that online support groups can provide a unique and valuable avenue through which healthcare professionals can learn more about the needs and experiences of patients.

In a recent study, Carroll, Koeling and Puri (2012) described experiments in the use of distributional similarity for acquiring lexical information from notes typed by primary care physicians who were general practitioners. They also present a novel approach to lexical acquisition from “sensitive” text, which does not require the text to be manually anonymized. This enables the use of much larger datasets compared to the situation where the sentences need to be manually anonymized and large datasets cannot be examined.

There is a considerable body of work that compares the practices of two popular social networking sites (Facebook & MySpace) related to trust and privacy concerns of their users, as well as self-disclosure of personal information and the development of new relationships (Dwyer, Hiltz & Passerini 2007). Scanfled, Scanfled and Larson (2010) studied the dissemination of health information through social networks. The authors reviewed Twitter status updates mentioning antibiotic(s) to determine overarching categories and explore evidence of misunderstanding or misuse of antibiotics. Most of the work uses only in-house lists of medical terms (Sokolova & Schramm 2011), each built for specific purposes, but they do not use existing electronic resources of medical terms designed for analysis of text from biomedical domain. However, those resources are general and in need of evaluation with respect to their applicability to the PHI extraction from social networks. The presented work fills this gap.

In most of the above cases, the authors analyze text manually and do not use automated text analysis. In contrast, in this work which we describe below, we want to develop an automated method for mining and analysis of personal health information.

5.3 Technology

5.3.1 Data mining

Data mining consists of building models to detect the patterns residing in data which allows us to classify and categorize data and extract more information that is buried in the data (Witten, Frank & Hall 2011). For example, find the co-relation between breast cancer and genetics of the patients, search for customers who are more interested in buying a product, and so forth. In order to analyze a problem, data mining extracts previous information and uses it to find solutions to the problem.

Data mining analyzes data stored in data repositories and databases. For example, it is possible that data originates from a business environment that

deals with information regarding a product, which helps management determine different market strategies such as which customers are more interested to buy a product based on their shopping history. This data can be used to contrast and compare different enterprises. Using data mining, a company can have real-time analysis of their day-to-day business activity, advertisement, promotion and sales purposes, as well as competing with the rival companies.

5.3.2 Machine learning

Machine learning can be considered as science for the design of computational methods using experience to improve their performance (Mitchell 1997). The algorithms in machine learning on large-scale problems, make accurate predictions, and address a variety of learning problems (e.g., classification, aggression, clustering of data). Examples of such methods are: K-Nearest Neighbor (e.g., What are the K members of a group that have similar shopping habits, time of shopping and items that they purchase), Naïve Bayes (e.g., classify whether if a certain product will be purchased or not, based on its price, brand name and so on), association rule mining (e.g., from lists of purchases in a grocery store, find which items are bought together by the customers, e.g., most of the customers that buy milk, buy cookies).

We use machine learning in many domains including the following examples:

- Text: Text and document classification, spam detection, morphological analysis, statistical parsing, fraud detection (credit card, telephone);
- Audio/Visual: Optical character recognition, part-of-speech tagging, speech recognition, speech synthesis, speaker verification, image recognition, face recognition;
- Other: Network intrusion, games, unassisted control of a vehicle (robots, navigation), medical diagnosis and many more.

In general, machine learning answers the questions such as: What can be learned, and under what conditions? What learning guarantees can be given? What is the algorithmic complexity?

5.3.3 Information extraction

Extracting structured information from an unstructured or semi-structured machine-readable document is referred to as information extraction (IE), which

is mostly done on natural language texts (Mitkov 2003). Examples of information extraction from social media can be discovering the general public opinion about a movie or a president's speech through comments and posts on social networks and weblogs.

The emergence of the user written web content created a greater need for the development of IE systems to assist people in dealing with the large amount of online data (Sokolova & Lapalme 2011). These systems should be scalable, flexible, and efficiently maintained.

5.3.4 Natural language processing

In computer science natural language processing (NLP) is referred to as a subsection of artificial intelligence and linguistics that addresses the interactions between computers and natural (human) languages. One of the main challenges in NLP is to enable computers to derive and understand meaning expressed by humans in text.

Modern NLP algorithms are based on machine learning methods. The idea of using machine learning is different from previous approaches to language processing, which involved direct hand coding of large sets of rules. That is, machine learning requires general learning algorithms to automatically learn such rules through the analysis of large input data also known as corpus. For example, in a sentence like *Boeing is located in Seattle*, a relationship between *Boeing*, tagged as a company, and *Seattle*, tagged as a location is derived and the system learns that this sentence is discussing a company-location relationship. These corpora are generally hand-annotated with the correct values to be learned (Jurafsky and Martin 2008), e.g., British National Corpus (BNC)² and Penn Treebank³.

For example, NLP tasks have used different classes of machine learning algorithms, which require the input of large set of "features" (e.g., sentence length, word count, punctuation and characters) generated from the corpus. Common "if-then" rules were used in the algorithms, such as decision trees. Gradually, research has tended to focus more on models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. These models can express more than one possible solution, which are more flexible and reliable (Manning & Schütze 2003).

² <http://www.natcorp.ox.ac.uk/>

³ <http://www.cis.upenn.edu/~treebank/>

5.4 Electronic resources of medical terminology

Biomedical information extraction and text classification have a successful history of method and tool development, including deployed information retrieval systems, knowledge resources and ontologies (Yu 2006). However, these resources are designed to analyze knowledge-rich biomedical literature. For example, GENIA is built for the microbiology domain. Its categories include DNA-metabolism, protein metabolism, and cellular process. Another resource, Medical Subjects Heading (MeSH), is a controlled vocabulary thesaurus, whose terms are informative to experts but might not be used by the general public. The Medical Entities Dictionary (MED) is an ontology containing approximately 60,000 concepts, 208,000 synonyms, and 84,000 hierarchies. Table 5.1 shows a sample of the MedDRA hierarchy. It shows *Biliary disorders* as one of the main categories under which there are many sub-categories including *Biliary neoplasms*. Furthermore, *Biliary neoplasms* can have a sub-category called *Biliary neoplasms benign*. In this hierarchy each sub-category gives more detailed information than its super category. This powerful lexical and knowledge resource is designed with medical research vocabulary in mind. The Unified Medical Language System (UMLS) has 135 semantic types and 54 relations that include organisms, anatomical structures, biological functions, chemicals, etc. Specialized ontology BioCaster was built for surveillance of traditional media. It helps to find disease outbreaks and predict possible epidemic threats. All these sources would require considerable modification before they could be used for analysis of messages posted on public Web forums.

5.4.1 MedDRA and its use in text data mining

The Medical Dictionary for Regulatory Activities (MedDRA) is an international medical classification for medical terms and drugs terminology used by medical

Tab. 5.1: A sample of the MedDRA hierarchy and their labels.

Category label	Main category	First level sub-category	Second level sub-category
10	Biliary disorders		
10-1		Biliary neoplasms	
10-1-1			Biliary neoplasms benign

professionals and industries. The standard set of MedDRA terms enables these users to exchange and analyze their medical data in a unified way. MedDRA has a hierarchical structure with 83 main categories in which some have up to five levels of sub-categories. MedDRA contains more than 11,400 nodes which are instances of medical terms, symptoms, etc.

Since its appearance nearly a decade ago, MedDRA has been used by the research community to analyze the medical records provided or collected by healthcare professionals: e.g., McLernon et al. (2010) use MedDRA in their study to evaluate patient reporting of adverse drug reactions to the UK “Yellow Card Scheme.” In another study, Star et al. (2011) use MedDRA to group adverse reactions and drugs derived from reports were extracted from the World Health Organization (WHO) global ICSR database that originated from 97 countries from 1995 until February 2010.

The above examples show that the corpus on which MedDRA tested is generally derived from patients’ medical history or other medical descriptions found in the structured medical documents that are collected or disclosed by healthcare professionals. Content and context of those documents differ considerably from those of messages written by social network users. In our study, we aim to evaluate the usefulness of MedDRA in detection of PHI disclosed on social networks.

5.4.2 SNOMED and its use in text data mining

Another internationally recognized classification scheme is the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) maintained by the International Health Terminology Standards Development Organization. Although SNOMED is considered the most comprehensive clinical healthcare terminology classification system, it is primarily used in standardization of electronic medical records (Campbell, Xu & Wah Fung 2011).

Medical terms in SNOMED are called *concepts*. A concept is indicative of a particular meaning. Each concept has a unique *id* that with which it is referred. A concept has a *description* which is a string used to represent a concept. It is used to explain what the concept is about. *Relationship* is a tuple of (*object – attribute – value*) connecting two concepts through an attribute.

Same as MedDRA, SNOMED has also a hierarchical structure. The root node, *SNOMED Concept*, has 19 direct children which Fig. 5.3 shows 10 of them from *Clinical finding* to *Record artifact*. As illustrated in Fig. 5.3, one of the nodes, *procedure*, has 27 branches including but not limited to, *administrative*

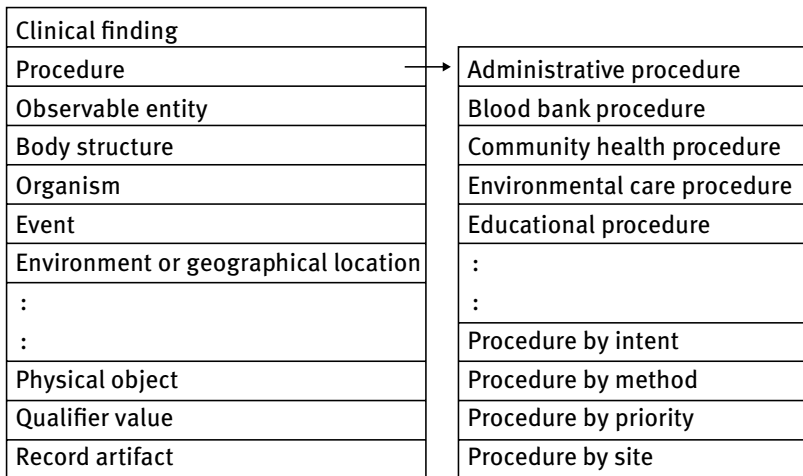


Fig. 5.3: A sample of the SNOMED hierarchy.

procedures (e.g., medical records transfer), *education procedures* (e.g., low salt diet education) and other procedures.

Among 353,154 instances of all 19 main branches we decided to only sub-select *procedures* and *clinical findings* (that encompasses *diseases* and *disorders*). These branches have more medical meanings than for instance the *Environment or geographical location* node which covers name of the cities, provinces/states, etc. The *clinical findings* node has 29,724 sub-nodes (19,349 *diseases and disorders*, 10,375 *findings*) and the node *procedure* has 15,078 sub-nodes. So in total we have selected 44,802 nodes out of 353,154.

Table 5.2 depicts a brief comparison between MedDRA and SNOMED hierarchical structure. It shows that SNOMED covers a larger set of terms and has deeper hierarchical levels compared to MedDRA.

Tab. 5.2: MedDRA and SNOMED hierarchical structure.

Dictionary	# Total nodes	# Unique sub-selected nodes	Average depth level
MedDRA	11,400	8561	3
SNOMED	353,154	44,802	6

5.4.3 Benefits of using MedDRA and SNOMED

We believe referring to MedDRA and SNOMED as well-funded, well-studied and reliable sources has two benefits:

1. We introduce a field of new text applications (the posts, weblogs and other information sources directly written by individuals) which extend the use for MedDRA and SNOMED. These medical dictionaries were previously used only for the health information collected by healthcare professionals.
2. Since MedDRA and SNOMED have well-formed hierarchical structures, by examining them against the posts on the social network site, we should be able to identify which terms and branches in the MedDRA and SNOMED are used to identify PHI and which branches are not, and thus can be pruned. These operations should result in a more concise and practical dictionary that can be used on detecting PHI disclosed in diverse textual environments.

5.5 Empirical study

In this research, we examined the amount of PHI disclosed by individuals on an online social network site, MySpace. Unlike previous research work, introduced in Section 2, the presence of PHI was detected through the use of the medical terminologies of MedDRA and SNOMED.

In our empirical studies, we examined posts and comments publicly available on MySpace. We sorted and categorized the terms used in both MedDRA and SNOMED, and found in MySpace, based on the frequency of their use and whether they reveal PHI or not. We also studied the hierarchy branches that are used and the possibility of pruning the unused branches (if any exists). Based on the hierarchical structures of MedDRA and SNOMED, the deeper we traversed down the branches, the more explicit the medical terms become and the harder the pruning phase is.

5.5.1 MySpace data

MySpace is an online social networking site that people can share their thoughts, photos and other information on their profile or general bulletin, i.e., posts posted on to a “bulletin board” for everyone on a MySpace user’s friends list to see. There have been several research publications on use of MySpace

data in text data mining, but none of them analyzed disclosure of personal health information in posted messages (Grace et al. 2007; Shani, Chickering & Meek 2008).

We obtained the MySpace data set from the repository of training and test data released by the workshop Content Analysis for the Web 2.0 (CAW 2009). The data creators stated that those datasets intended to comprehend a representative sample of what can be found in web 2.0. Our corpus was collected from more than 11,800 posts on MySpace. In the text pre-processing phase we eliminated numbers, prepositions and stop words. We also performed stemming which converted all the words to their stems (e.g., hospital, hospitals and hospitalized are treated the same).

5.5.2 Data annotation

We manually reviewed 11,800 posts on MySpace to see to what extent those medical terms are actually revealing personal health information on MySpace. The terms were categorized into three groups:

- PHI: terms revealing personal health information.
- HI: medical terms that address health information (but not necessarily personal).
- NHI: terms with non-medical meaning.

To clarify this let us see the following examples:

The word *lung* which assumed to be a medical term appears in the following three sentences we got from our MySpace corpus: “... they are promoting cancer awareness particularly lung cancer ...” which is a medical term but does not reveal any personal health information. “... I had a rare condition and half of my lung had to be removed...” this is clearly a privacy breach and “... I saw a guy chasing someone and screaming at the top of his lungs ...” which carries no medical value. In this manner, we have manually analyzed and performed manual labelling based on the annotator’s judgment whether the post reveals information about the person who wrote it, or discloses information about other individuals that make them identifiable.

We acknowledge that there might be cases where the person might be identified with a high probability in posts that mention “... my aunt ..., my roommate.” For simplicity in this research we categorize those posts as HI where the post has medical values but does not reveal a PHI. Table 5.3 shows some more examples of PHI, HI and NHI.

Tab. 5.3: Examples of terms found on MySpace which are PHI, HI and NHI.

Term	PHI	HI	NHI
Fraction	... got a huge bump on my forehead, fractured my nose	I wish the driver would've died as well instead of just suffering a fractured leg	The few people who did vote would be so fractured among the different parties
Laser	For me the laser treatment had unpleasant side effects	I know someone who had laser surgery to remove the hair from his chest	... with a laser writes something on a flower stem
Allergic	I'm allergic to cigarette smoke	... the allergy is a valid reason	I'm allergic to bullets!

5.5.3 MedDRA results

To assess MedDRA's usability for PHI detection, we performed two major steps:

1. We labeled the MedDRA hierarchy in a way that the label of each node reflects to which branch it belongs. The result is corpus-independent.
2. We did uni-gram and bi-gram (a contiguous sequence of one term from a given sequence of text or speech) comparisons between the terms that appear in MySpace and the words detected by MedDRA. The result is corpus-dependent.

After the execution of the step 1, MedDRA's main categories are labeled from 1 to 83 and for those with consequent sub-categories, the main category number is followed by a hyphen (-) and the sub-category's number [e.g., *Biliary disorders* (10) and its sub categories *Biliary neoplasms* (10-1) and *Biliary neoplasms benign* (10-1-1)].

After the execution of the step 2, there are 87 terms that appear both in MedDRA and in the MySpace corpus. A subset of them is illustrated in Tab. 5.4.

There are also identical terms that appear under different categories and increase the ambiguity of the term. For instance, *nausea* appears under categories *acute pancreatitis* and in *gastrointestinal nonspecific symptoms and therapeutic procedures*, so when *nausea* appears in a post, it is not initially clear which

Tab. 5.4: A subset of terms detected by MedDRA that appear in MySpace.

Terms	PHI	HI	NHI	Terms	PHI	HI	NHI
Depression	18	114	0	Dizzy	2	7	2
Injury	9	12	0	Overdose	2	6	0
Swell	4	2	1	Thyroid	2	0	0
Concussions	3	0	0	Asthma	1	1	0

category of the MedDRA hierarchy has been used, and the text needs further semantic processing.

5.5.4 SNOMED results

SNOMED leaves are very specific and have many more medical terms compared to MedDRA. Our manual analysis has shown that the general public uses less technical, and therefore more general, terms when they discuss personal health and medical conditions. Hence, we expect that SNOMED's less granular (defined as less specific in identifying information) terms appear more often in MySpace data than their more specific counterparts, which have higher granularity.

The structure of SNOMED is organized as follows: The root node has 19 sub-nodes. One of the sub-nodes is *procedure* that itself has 27 sub-nodes. One of those sub-nodes is called *procedure by method* which has 134 sub-nodes. *Counseling* is one of those 134 sub-nodes and itself has 123 sub-nodes. Another node among the 134 sub-nodes is *cardiac pacing* that has 12 sub-nodes which are mostly leaves of the hierarchy.

In extreme cases there might be nodes that are located 11 levels deep down the hierarchy. For example the following shows the hierarchy associated with the node *hermaphroditism*. Each “>” symbol can be interpreted as “is a ...”:

Hermaphroditism > Disorder of endocrine gonad > Disorder of reproductive system > Disorder of the genitourinary system > Disorder of pelvic region > Finding of pelvic structure > Finding of trunk structure > Finding of body region > Finding by site > Clinical finding > SNOMED Concept

It is cumbersome to understand how many of the 44,802 nodes that we have sub-selected from SNOMED are leaves and how many are intermediate nodes; however, 66 nodes out of 44,802 appeared in MySpace, of which nine were leaves (see Tab. 5.5).

Tab. 5.5: A subset of terms detected by SNOMED that appear in MySpace.

Terms	PHI	HI	NHI	Terms	PHI	HI	NHI
Sick	44	1	135	Fracture	3	3	1
Pain	17	3	141	Dizzy	2	7	2
Infection	5	33	0	Insomnia	2	6	0
Swell	4	2	1	Thyroid	2	0	0

Table 5.6 shows some terms. The number of times the term appears in SNOMED, and whether it is a leaf in the hierarchy. We can see that except *jet lag* and *dizzy*, the other terms do not reveal PHI. Even in the example *dizzy*, the appearance as PHI compared to the number of times they appear as HI and NHI is trivial.

This result indicates that although SNOMED has a deep hierarchical structure, one should not traverse all the nodes and branches to reach leaf nodes to be able to detect PHI terms. In contrast, we hypothesize that branches can be pruned to reduce the PHI detection time and still achieve an acceptable result. We leave this as potential future work.

5.6 Risk factor of personal information

5.6.1 Introducing RFPI

Due to the semantic ambiguity of the terms we had to manually examine the given context to see whether the terms were used for describing medical concepts or not. For instance, the term *adult* in the post “... today young people indifferent to the adult world ...” has no medical meaning.

We aimed to find whether the terms were used for revealing PHI or HI. Although some terms like *surgery* and *asthma* have strictly (or with high certainty) medical meaning, some terms may convey different meanings depending on where or how they are used. For instance, the word *heart* has two different meanings in “... heart attack ...” and “... follow your own heart ...”.

Tab. 5.6: Terms of MySpace detected by SNOMED leaf nodes.

Term	Level of hierarchy	PHI	HI	NHI	# Freq. in SNOMED
Phlebotomy	3	0	3	0	2
Histology	4	0	2	0	2
Jet lag	4	1	0	0	1
Domestic abuse	5	0	1	0	1
Physic assault/abuse	5	0	5	0	1
Black out	6	0	1	1	1
Dizzy	6	2	7	2	2
Hematological	6	0	1	0	1
Papillary conjunctivitis	6	0	1	0	2

We also measured the ratio of the number of times that the term was used in MySpace and the number of times that revealed PHI. We called the ratio the Risk Factor of Personal Information (RFPI). In other words, for a term t , $RFPI_t$ is:

$$RFPI_t = \text{number of times } t \text{ reveals PHI} / \text{number of times } t \text{ appears in a text}$$

Table 5.7 illustrates the top RFPI terms from MedDRA and SNOMED that often reveal PHI. There is an overlap between the top most used terms of MedDRA and SNOMED with highest RFPI. These are terms that prone to the number of times they appear in data (*concussions, thyroid, hypothermia, swell, ulcer and fracture*).

Furthermore, according to our studies although the words *sick* and *pain* appear numerous times and reveal personal health information their RFPI is relatively low and might not be as privacy-revealing as words like *fracture* or *thyroid*.

For example *sick* in the sentence “... I am sick and tired of your attitude ...” or “... the way people were treated made me sick ...” clearly belong to the NHI group and does not carry any medical information. Or in the case of the term *pain*, the sentences “... having a high-school next to your house is going to be a pain ...” or “... I totally feel your pain! ...” belong to NHI group as well.

5.6.2 Results from MedDRA and SNOMED

In total, we found 127 terms that appear in MySpace and in both dictionaries. 87 terms in MySpace are captured by MedDRA and 66 terms are captured by

Tab. 5.7: Top terms detected by MedDRA and SNOMED that have highest RFPI.

a) MedDRA					b) SNOMED				
Term	PHI	HI	NHI	RFPI	Term	PHI	HI	NHI	RFPI
Concussions	3	0	0	1.00	Concussions	3	0	0	1.00
Thyroid	2	0	0	1.00	Thyroid	2	0	0	1.00
Disoriented	1	0	0	1.00	Bipolar disorder	1	0	0	1.00
Hyperthyroid	1	0	0	1.00	Hypothermia	1	0	0	1.00
Hypothermia	1	0	0	1.00	Jet lag	1	0	0	1.00
Liposuction	1	0	0	1.00	Motion sickness	1	0	0	1.00
Swell	4	2	1	0.57	Shoulder pain	1	0	0	1.00
Asthma	1	1	0	0.50	Stab wound	1	0	0	1.00
Ulcer	1	1	0	0.50	Tetanus	1	0	0	1.00
Injury	9	12	0	0.43	Thyrotoxicosis	1	0	0	1.00
Fracture	3	3	1	0.43	Swell	4	2	1	0.57

SNOMED. There are 26 common terms that appear in both dictionaries. Although SNOMED is a larger dataset compared to MedDRA, since its terms are more specific, fewer terms are appeared in SNOMED. Thus, we consider MedDRA to be more useful for PHI detection.

Figure 5.4 illustrates the number of sentences (not terms) in MySpace for each category of PHI, HI and NHI that are detected by MedDRA, SNOMED and the union of them. Table 5.8 demonstrates that although SNOMED detects more PHI terms compared to MedDRA, since it also detects more NHI terms (false positive) as well, it is less useful compared to MedDRA. In addition, the summation of both PHI and HI in MedDRA is greater than its equivalent in SNOMED which is another reason why MedDRA seems more useful than SNOMED.

In brief, since MedDRA covers a broader and more general area it detects more HI than SNOMED. In contrast, although SNOMED detects slightly more PHI and HI, it is less trustable than MedDRA since it also detects far more NHI. Hence, the precision of detection is much lower.

Although there are not many sentences (nine sentences out of almost 1000 sentences) that reveal PHI as a result of engaging in a conversation that initially contained HI, there is always the possibility that existence of HI sentences is more likely to result in PHI detection compared to the sentences that contain NHI.

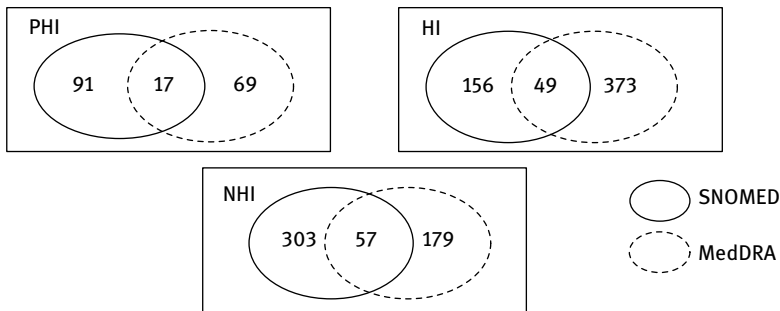


Fig. 5.4: HI, PHI and NHI sentences detected by each dictionary and their intersection.

Tab. 5.8: Percentage of the sentences that are detected by these two dictionaries in each group.

Dictionary	%PHI	%HI	%NHI
SNOMED	16.5	28.5	55
MEDDRA	11.5	60	28.5

As shown in Fig. 5.4, the amount of terms that are detected by both MedDRA and SNOMED (their intersection in the Venn diagram) is not impressive and that is why these two dictionaries cannot be used interchangeably.

5.6.3 Challenges in detecting PHI

We wanted to estimate how many PHI sentences our method could miss. To make a rigorous estimate we decided on a *manual* evaluation. For this purpose, we used the Linux command:

```
shuf -n input | head -n 1000 > output
```

We tested the command by repeating it three times to make sure each time it produces a completely random list of 1000 comments in the corpus. On the fourth try, we extracted 1000 sentences and manually assessed them. We have seen that our algorithm missed only the following three PHI comments:

- “ok i’m off to bed before this vicadine wears off yay head spinning sleep.
- never come back i’m bored i took some vicadin so i should be goin to sleep soon. yay!
- nope! 2 blurry of an photo. she may have an cold sore on her lips.”

In the first two comments the name of the drug was misspelled (the correct spelling is Vicodin) and could not be detected by either MedDRA or SNOMED.

Regarding the third sentence, PHI can be expressed in a descriptive way. Functionally, in PHI description, some words cannot be modified, whereas associated words can be changed (Sokolova & Lapalme 2011). Those words are called descriptors. For instance, in *hot water*, *boiling water* and *cold water*, the word *water* represents the target concept, thus cannot be modified. The accompanying words *hot*, *boiling*, and *cold* can change depending on the context. In our case, the word *sore* in MedDRA is used as a modifier, such as *sore throat*, whereas in the term *cold sore* it is used as a descriptor. Hence, in the third sentence, the term *cold sore* was used that could not be detected by our medical dictionaries.

We found three sentences in one thousand comments that contained some sort of PHI and were not detected by our algorithm. Our algorithm gives very impressive result in terms of sentences that it detects that have PHI and only an estimated false negative of 0.003 which is considered a great result for missing PHI compared to the state of the art false negative rate of 2.9–3.9 (Miles, Rodrigues & Sevdalis 2013). The obtained number missed PHI enhances our previous results reported in (Ghazinour, Sokolova & Matwin 2013).

5.7 Learning the profile of PHI disclosure

We approach the task of detecting PHI leaks as acquiring a profile of what “language” is characteristic of this phenomenon which occurs in posts on health-related social networks. This can be achieved if a profile of the occurrence of this phenomenon is acquired. A machine learning, or more specifically text classification, is a natural technique to perform this acquisition of a profile. We studied the classification of the sentences under two categories of PHI-HI and NHI using Machine Learning methods.

Hypothesis. Focusing on terms from MedDRA and SNOMERD results in a better performing profiling than the straightforward method of a bag of words.

Experiment. Our experiment consists of the following two parts:

5.7.1 Part I – Standard bag of words model

We vectorized each of the 976 sentence detected by the two medical dictionaries. In these sentences, there are 1865 distinct terms. After removing the words with the same roots and deleting the symbols and numerical terms, 1669 unique words were identified. Next, generating a standard Bag of Words document representation and the sentences are vectorized (0 for not existing and 1 for existing). Hence, we have vectors of 1669 attributes that are either 0 or 1 and one more attribute which is the label of the sentence, the privacy class (0 = NHI, 1 = PHI-HI).

After each vector is labeled accordingly to be either PHI-HI or NHI, we perform a bi-classification and train our model with 976 sentences of which 425 are labeled as NHI and 551 are labeled as PHI-HI.

We used two classification methods used most often in text classification, i.e., Naive Bayes (NB), KNN (IBK) in Weka based on the privacy class (0 = NHI, 1 = PHI-HI) shown in the left column of Tab. 5.9. Hence, our training data set would be the sentences with binary values of the terms appearing in them or not. Due to our small set data, we performed five by two cross-validations. In each fold, our collected data set were randomly partitioned into two equal-sized sets in which one was the training set which was tested on the other set. Then we calculated the average and variance of those five iterations for the privacy class. Table 9 shows the results of this 5×2-fold cross validations.

Tab. 5.9: Two classification methods on the privacy class with and without medical terms.

Classification		Privacy class (Part I)	Privacy class (Part II)
NB	Correctly classified%	75.51	85.75
	Mean absolute error	0.25	0.15
KNN (k=2)	Correctly classified%	74.48	86.88
	Mean absolute error	0.27	0.13

5.7.2 Part II – Special treatment for medical terms

We took the vectors resulting from Part I and focused on the terms belong to the following three groups by weighting them stronger in the bag of words than the remaining words.

- a) List of pronouns or possessive pronouns, members/relatives (e.g., I, my, his, her, their, brother, sister, father, mother, spouse, wife, husband, ex-husband, partner, boyfriend, girlfriend, etc.).
- b) Medical term detected by MedDRA and SNOMED.
- c) Other medical terms that their existence in a sentence may result in a sentence to be a PHI or HI. Terms such as *hospital*, *clinic*, *insurance*, *surgery*, etc.

For group (a) and group (c) we associate weight 2 (one level more than the regular terms that are presented by 1 as an indication that the terms exist in the sentence). For group (b) which are the terms detected by the SNOMED and MEDRA and have higher value for us we associate weight with value of 3. So unlike the vectors in Part I which consist of 0s or 1s, in this part we have vectors of 0s, 1s, 2s and 3s. In fact, the values for weights are arbitrary and finding the right weights would be the task of optimization of the risk factor. We ran the experiment with values of 0s, 1s, 2s and 4s and we got the same results shown in Tab. 5.9.

Next, we used the same two classification methods and performed five by two cross-validations. The results are shown in Tab. 5.9 in the right column. Comparing the results from Part I and II show that there is an almost 10% improvement in detecting sentences that reveal health information using the terms detected by MedDRA and SNOMED which confirms our hypothesis. The results are statistically significant (Dietterich 1998) with the *p-value* of 0.95.

5.8 Conclusion and future work

In this research work, we studied personal health information (PHI) by means of data mining, machine learning, natural language processing and information

extraction. We showed that such models can detect and, when necessary, protect personal health information that might be unknowingly revealed by users of social networks. In this work we presented empirical support for this hypothesis, and furthermore we showed how two existing electronic medical resources MedDRA and SNOMED helped to detect personal health information in messages retrieved from a social network. Our algorithm gives robust results in terms of sentences and phrases that it detects as containing PHI and only an estimated false negative of 0.003. This is considered a valid and reliable result for missing PHI compared to the state of the art false negative rate of 2.9–3.9 (Miles, Rodrigues & Sevdalis 2013). Our current work enhances the results reported in (Ghazinour, Sokolova & Matwin 2013).

We showed how existing medical dictionaries can be used to identify PHI. We labeled the MedDRA and SNOMED hierarchy in a way that the label of each node reflects which branch it belongs. Next, we did uni-gram and bi-gram comparisons between the terms that appear on the MySpace corpus and the words that appear on MedDRA and SNOMED. Comparing the number of terms captured by these two medical dictionaries, it suggests that MedDRA covers more general terms and seems more useful than SNOMED that has more detailed and descriptive nodes. Performing a bi-classification on the vectors resulted from the sentences labeled as PHI-HI and NHI support our hypotheses. We used two common classification methods to validate our hypothesis and analyse the results of MedDRA and SNOMED. Our experiments demonstrated that using the terms detected by MedDRA and SNOMED helps us to better identify sentences in which people reveal health information.

Future directions include analysis of words which tend to correlate but not perfectly match the terms contained in medical dictionaries (e.g., in the sentence “I had my bell rung in the hockey game last night”: the phrase “bell rung” which means a nasty blow suffered to the head during a sports game that may indicate a concussion). Researchers can use methods such as Latent Dirichlet Allocation (LDA) in such instances. Furthermore, testing our model (which we based on the MySpace social media network) on different posts found on other social networks, such as Facebook or Twitter, would be a good research experiment. We would want to compare the terms that appear in MedDRA and SNOMED and evaluate their RFPI values in analyzing the health-related posts of these other social networks.

An interesting project would be to develop a user interface or an application that could be plugged into the current social networks such as MySpace, Facebook and Twitter or appear on the user’s smartphone or tablet that would essentially warn the user about revealing PHI when they use these potentially privacy-violating words that we have introduced in this research.

Another potential future work could investigate use of more advanced NLP tools, beyond the lexical level, to identify some of the semantic structures that contain terms that might lead to health-related privacy violations.

Acknowledgment

The authors thank NSERC for the funding of the project and anonymous reviewers for many helpful comments.

References

- Balocco, L. & Paganelli, C. (2011) *Access to Health Information: Going From Professional to Public Practices Information Systems and Economic Intelligence*: 4th Int. Conference - SII'E'2011.
- Bansal, G., Zahedi, F. M. & Gefen, D. (2010) 'The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online', *Decis Support Syst*, 49(2):138–150.
- Bobicev, V., Sokolova, M., Jafer, Y. & Schramm, D. (2012) *Learning Sentiments from Tweets with Personal Health Information*. Canadian Conference on AI, pp. 37–48.
- Campbell, J., Xu, J. & Wah Fung, K. (2011) 'Can SNOMED CT fulfill the vision of a compositional terminology? Analyzing the use case for Problem List', *AMIA Annu Sympos Proc*, 181–188.
- Carroll, J., Koeling, R. & Puri, S. (2012) 'Lexical acquisition for clinical text mining using distributional similarity. Computational linguistics and intelligent text processing', *Lect Notes Comput Sci*, 7182:232–246.
- Dietterich, T. G. (1998) 'Approximate statistical tests for comparing supervised classification learning algorithms', *Neural Computn*, 10(7):1895–1924.
- Dwyer, C., Hiltz, S. R. & Passerini, K. (2007) 'Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace'. *Proceedings of the Thirteenth Americas Conference on Information Systems*, Keystone, Colorado, August.
- Fox, S. (2011) *The Social Life of Health Information*. Pew Research Center's Internet & American Life Project. [online] at: http://www.pewinternet.org/~media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf [20 October 2013].
- Ghazinour, K., Sokolova, M. & Matwin, S. (2013) 'Detecting Health-Related Privacy Leaks in Social Networks Using Text Mining Tools', *Canadian Conference on AI*, pp. 25–39.
- Grace, J., Gruhl, D., Haas, K., Nagarajan, M., Robson, C. & Sahoo, N. (2007) *Artist ranking through analysis of on-line community comments* from: [http://domino.research.ibm.com/library/cyberdig.nsf/papers/E50790E50756F371154852573870068A371154852573870184/\\$File/rj371154852573810421.pdf](http://domino.research.ibm.com/library/cyberdig.nsf/papers/E50790E50756F371154852573870068A371154852573870184/$File/rj371154852573810421.pdf).
- Gross, R. & Acquisti, A. (2005) 'Information revelation and privacy in online social networks (the facebook case)', In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pp. 71–80.

- Harries, D. & Yellowlees, P. M. (2013) 'Cyberterrorism: Is the U.S. healthcare system safe?', *Telemed e-Health* 19(1):61–66.
- Jurafsky, D. & Martin, J. H. (2008) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Second Edition, Prentice-Hall, Upper Saddle River, NJ.
- Keckley, P. H. & Hoffman, M. (2010) *Social Networks in Health Care: Communication, Collaboration and Insights*, Deloitte Development LLC.
- Kennedy, D. (2012) 'Doctor blogs raise concerns about patient privacy'. Available at: www.npr.org/templates/story/story.php?storyId#equal#88163567. [June 13, 2012].
- Knudson, J. (2013) 'Healthcare information, the new terrorist target', *The Record*, 25(6):10.
- Lagu, T., Kaufman, E., Asch, D. & Armstrong, K. (2008) 'Content of weblogs written by health professionals', *J Gen Intern Med*, 23(10):1642–1646.
- Li, F., Zou, X., Liu, P. & Chan, J. Y. (2011) 'New threats to health data privacy', *BMC Bioinformatics*, 12:57.
- Malik, S. & Coulson, N. (2010) 'Coping with infertility online: an examination of self-help mechanisms in an online infertility support group', *Patient Educ Couns*, 81:315–318.
- Manning, C. D. & Schütze, H. (2003) *Foundations of Statistical Natural Language Processing*, MIT Press: Cambridge, MA, Sixth Printing.
- McLernon, D. J., Bond, C. M., Hannaford, P. C., Watson, M. C., Lee, A. J., Hazell, L. & Avery, A. (2010) 'Adverse drug reaction reporting in the UK: a retrospective observational comparison of Yellow Card reports submitted by patients and healthcare professionals', *Drug Safety*, 33(9):775–788.
- MedDRA Maintenance and Support Services Organization. [online] <http://www.meddrasso.com>. [Jan 1, 2013].
- Miles, A., Rodrigues, V. & Sevdalis, N. (2013) 'The effect of information about false negative and false positive rates on people's attitudes towards colorectal cancer screening using faecal occult blood testing (FOBT)', *Patient Educ Couns*, 93(2):342–349.
- Miller, A. R. & Tucker C. (2009) 'Privacy protection and technology adoption: The case of electronic medical records', *Manag Sci*, 55(7):1077–1093.
- Mitchell, T. M. (1997) *Machine learning*. New York, NY: The McGraw-Hill Companies, Inc.
- Mitkov, R. (2003) *The Oxford Handbook of Computational Linguistics*. Oxford University Press.
- Princeton University (2010) *About WordNet*. WordNet. Princeton University, [online] Available at: <http://wordnet.princeton.edu>. [8th October, 2013].
- Renahy, E. (2008) *Recherche bd'infomation en matiere de sante sur INternet: determinants, pratiques et impact sur la sante et le recours aux soins*, Paris 6.
- Scanfeld, D., Scanfeld, V. & Larson, E. (2010) 'Dissemination of health information through social networks: twitter and antibiotics', *American Journal of Infection Control*, 38(3):182–188.
- Shani, G., Chickering, D. M. & Meek, C. (2008) 'Mining recommendations from the web'. In RecSys '08: Proceedings of the 2008 ACM Conference on Recommender Systems, pp. 35–42.
- Silverman, E. (2008) 'Doctor Blogs Reveal Patient Info & Endorse Products'. Pharnalot www.pharnalot.com/2008/07/doctor-blogs-reveal-patient-info-endorse-products/. [Dec. 15, 2009].
- Sokolova, M. & Lapalme, G. (2011) 'Learning opinions in user-generated web content', *Nat Lang Eng*, 17(4):541–567.

- Sokolova, M., Matwin, S., Jafer, Y. & Schramm, D. (2013) 'How Joe and Jane Tweet about Their Health: Mining for Personal Health Information on Twitter'. In *the Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria, pp. 626–632.
- Sokolova, M. & Schramm, D. (2011) 'Building a patient-based ontology for mining user-written content'. In *Recent Advances in Natural Language, Processing Hissar, Bulgaria*, pp. 758–763.
- Star, K., Norén, G. N., Nordin, K. & Edwards, I. R. (2011) 'Suspected adverse drug reaction reported for children worldwide: an exploratory study using VigiBase', *Drug Safety*, 34:415–428.
- Systematized Nomenclature of Medicine. [online] www.ihstdo.org/snomed-ct/, [Jan 1, 2013].
- Witten, I. H., Frank, E. & Hall, M. A. (2011) *Data Mining: Practical Machine Learning Tools and Techniques*, (third edition) Morgan Kaufmann: Burlington, MA.
- Yeniterzi, R., Aberdeen, J., Bayer, S., Wellner, B., Clark, C., Hirschman, L. & Malin, B. (2010) 'Effects of personal identifier resynthesis on clinical text de-identification', *J Am Med Inform Assoc*, 17(2):159–168.
- Yu, F. (2006) *High Speed Deep Packet Inspection with Hardware Support*, Technical Report No. [online] UCB/EECS-2006-156; www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-156.html, [January 2013].
- Zhang, W., Gunter, C. A., Liebovitz, D., Tian, J. & Malin, B. (2011) 'Role prediction using electronic medical record system audits'. In *AMIA (American Medical Informatics Association) Annual Symposium*, pp. 858–867.

Hanna Suominen, Leif Hanlen and Cécile Paris

6 Twitter for health – building a social media search engine to better understand and curate laypersons’ personal experiences

Abstract: Healthcare professionals, trainees, and laypersons increasingly use social media over the Internet. As a result, the value of such platforms as a vital source of health information is widely acknowledged. These technologies bring a new dimension to health care by offering a communication medium for patients and professionals to interact, share, and survey information as well as support each other emotionally during an illness. Such active online discussions may also help in realizing the collective goal of improving healthcare outcomes and policies. However, in spite of the advantages of using social media as a vital communication medium for those seeking health information and for those studying social trends based on patient blog postings, this new medium of digital communication has its limitations too. Namely, the current inability to access and curate relevant information in the ever-increasing gamut of messages. In this chapter, we are seeking to understand and curate laypersons’ personal experiences on Twitter. To do so, we propose some solutions to improve search, summarization, and visualization capabilities for Twitter (or social media in general), in both real time and retrospectively. In essence, we provide a basic recipe for building a search engine for social media and then make it increasingly more intelligent through smarter processing and personalization of search queries, tweet messages, and search results. In addition, we address the summarization aspect by visualizing topical clusters in tweets and further classifying the retrieval results into topical categories that serve professionals in their work. Finally, we discuss information curation by automating the classification of the information sources as well as combining, comparing, and correlating tweets with other sources of health information. In discussing all these important features of social media search engines, we present systems, which we ourselves have developed that help to identify useful information in social media.

6.1 Introduction

Social media refers to interactions among people in which they create, share, exchange, or comment on information or ideas in the Internet or with other virtual communities and networks that predate the Web, such as bulletin board

services (Ahlqvist et al. 2008). Examples of social media include *Facebook*, *MySpace*, *PatientsLikeMe*, *Second Life*, *Twitter*, *Wikipedia*, and *YouTube*.¹

In 2013, Twitter is one of the most popular platforms for social media in the Internet. It was created and opened for users in March–June 2006. This platform enables users to create, share, exchange, and comment short messages (a.k.a. microblogs), called *tweets*. Each tweet has the maximum length of 140 characters and is connected with optional *metadata* for additional information, deeper context, and embedded media. On its seventh birthday in March 2013, Twitter had over 200 million active users creating over 400 million tweets per day (Twitter Blog 2013). To illustrate its rapidly increasing popularity, the numbers of active users and tweets per day were as follows: In the end of 2008, barely two million tweets were sent per day, and just a year and half later, in June 2010, the number of tweets per day was 65 million (Twitter Blog 2011). In June 2011, there were 200 million tweets daily, with 140 million active users, and, in March 2012, with approximately the same number of active users, volume had reached 340 million tweets per day (Twitter Blog 2011, 2012).

The value of social media as a source of health information has been widely acknowledged as useful to both healthcare professionals and *laypersons* alike. By following the definition of the Oxford dictionaries,² we use *laypersons* in this chapter as a reference to those without professional or specialized knowledge in health care. Based on the review of 98 studies by Moorhead et al. (2013), social media brings a *new dimension to health care* by offering a medium for *laypersons* and healthcare professionals to communicate about health issues with the potential of improving healthcare outcomes and policies. This new kind of online communication is seen to have many consequences. By increasing interactions between *laypersons* and professionals, vital health information as well as other users' subjective health experiences are made available to patients, their caretakers, and professionals who are monitoring and participating in these online chats. Much peer/social/emotional support is often derived from such communications. Moreover, the sheer volume of social media communications provides a means for facilitating a much improved public health surveillance. This new media is catching on among professionals. A review of 96 studies by Hamm et al. (2013) shows that the use of social media is already widespread among healthcare professionals and trainees, who see this networking medium as instrumental in helping to facilitate communication and in increasing one's knowledge. Patients likewise increasingly use the Internet and social media to discover knowledge and

1 Facebook (<https://www.facebook.com/>), MySpace (<https://myspace.com/>), PatientsLikeMe (<http://www.patientslikeme.com/>), Second Life (<http://secondlife.com/>), Twitter (<https://twitter.com/>), Wikipedia (<http://en.wikipedia.org/wiki/Wikipedia>), and YouTube (<http://www.youtube.com/>)

2 The Oxford dictionaries (<http://www.oxforddictionaries.com/>)

trends related to a variety of health problems, as well as for healthcare access and proper treatment (Young & Bloor 2009). For example, PatientsLikeMe, founded in 2004, reports now on their website having more than 220,000 users and covers more than 2000 medical conditions in 2013.

However, using social media as a source of health information has its limitations too. The primary limitation consists of the current inability to select, organize, and present – that is, *curate* – information on social media in terms of its quality and reliability (Moorhead et al. 2013). For example, when Hurricane Sandy killed more than 280 people and caused a total of nearly 70 billion dollars in damages in the central and north America in October 2012, as little as 9% of the related tweets were assessed as useful and reliable (The Australian 2013). This finding is based on analyzing a dataset of 50 million tweets over a 28-day period before, during, and after the hurricane. The second limitation is actually a *byproduct* of the popularity of social media itself. That is, the popularity of social media can sometimes make it hard to satisfy information seekers' needs for finding relevant and useful information because they must sift through a tremendous amount of messages found on the social media sites, blogs, bulletin boards, and other virtual communities and networks. This phenomenon of one's inability to sift through a voluminous amount of social media messages is a problem in *information access*.³ In sum, these two limitations – difficulties in curating the existing information and difficulties in sifting through mounds of social media posts – emphasize the need for new methods and approaches that can be used in identifying accurate and useful information from the multitude of tweets.

In this chapter, we are seeking to understand and curate laypersons' personal health experiences expressed as Twitter or other social media messages. We begin by detailing different sources of health information, positioning social media as one of those sources, illustrating the amount of information across such sources, and discussing the potential and limitations of social media as a source of health information. Then, we describe some solutions that have been proposed to improve Twitter search engines, either by an improved understanding of laypersons' experiences or by adding new ways of monitoring information quality and reliability. We provide a basic recipe for building a search engine for social media and methods for making it increasingly intelligent. This includes intelligent techniques to enrich the search queries; analyze the relevance of a given tweet and query; classify tweets into topical categories; curate their content through automated source classification; and combine, compare, and correlate tweets with

³ Information access is defined as satisfying people's information needs through natural, efficient interactions with an automated system that leverages world-wide structured and unstructured data in any language (Allan et al. 2003).

other sources of health information. We present systems we have developed that help look for useful information in social media. Finally, we discuss our research findings and conclusions.

6.2 Background

6.2.1 Social media as a source of health information

Today, social media forms a part of health information. Health information includes all health-related content in all data formats, document types, information systems, publication media, and languages from all specialties, organizations, regions, states, and countries across the dimensions of *audience* (e.g., clinician or layperson), *accessibility* (e.g., public or limited), and *accreditation* (e.g., official and approved, peer reviewed, or uncurated) (Fig. 6.1). This information also depends on the *author*, *community*, and *time* (Suominen et al. 2008). So,

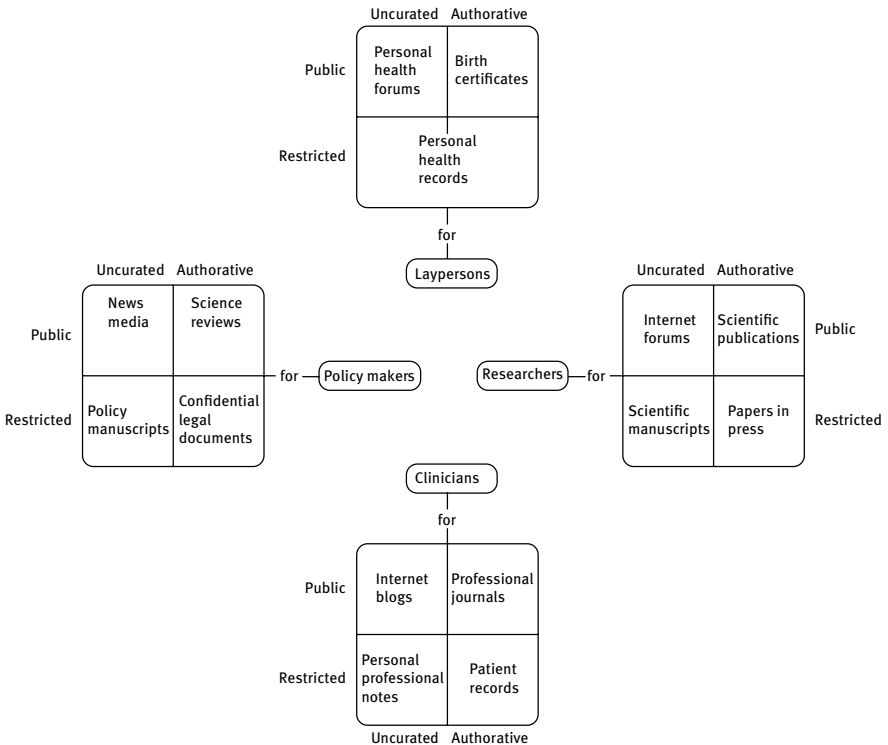


Fig. 6.1: Dimensions of health information.

for example, we all have our individual writing style, which evolves over time upon interacting with people at our home, school, and work. This style includes, among other things, our own particular use of jargon, abbreviations, and acronyms. Because we all have our own writing style, it is not uncommon to find tweets that, although addressing the very same topic, show many variations in vocabulary, style, and hashtags. This varying language is sometimes unintentional (e.g., misspellings or grammatical errors). Sometimes, however, such variations are by choice, thereby expressing the creativity of the individual tweeter.

Let us briefly illustrate the wide gamut of health information across these various dimensions indicated in Fig. 6.1. *PubMed*,⁴ which is a search engine accessing the MEDLINE database maintained by the United States National Library of Medicine, gives us an example of publicly available, peer reviewed scientific papers for researchers and practitioners. It is one of the most popular search engines in the Internet for biomedical literature, life science journals, and online books. Approximately 10 years ago in 2003, the MEDLINE database consisted of nearly 16 million entries (i.e., publications⁵). This number grew by almost 600,000 entries since 2002, that is, by more than 1600 entries per day. When we go forward 10 years to 2013, we see that the total number of PubMed entries has grown to over 23 million, and the respective growth rate has almost doubled to 3000 entries per day. However, if we do not limit ourselves to PubMed exclusively but look instead for *all* publications in health sciences appearing in the Internet, we see that, as far back as 2003, the daily growth rate of added entries exceeded 3300 papers (Coiera 2003). What this shows us is that the availability of health-related publications on the web has been growing exponentially over the years.

The Australian and US portals for clinical practice guidelines by (1) the Australian Government and National Health and Medical Research Council, and (2) the US Department of Health and Human Services, Agency for Healthcare Research Quality, respectively, are also curated and publicly available.⁶ Their over 4000 approved up-to-date documents are not only targeted to professionals but to laypersons as well.

In contrast, records produced in health care for professionals (and laypersons) are confidential and official. Within these records, the largest quantities

⁴ PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>)

⁵ We retrieved these statistics by comparing the number of returned entries on PubMed for the query of `1800:2003[dp]` (`2003:2003[dp]`) [`1800:2013[dp]`] [`2013:2013[dp]`] where `dp` refers to the date of publication and `1800:2003`, for example, to the date range from 1 January 1800 to 31 December 2003.

⁶ <http://www.clinicalguidelines.gov.au/> by the Australian Government, National Health and Medical Research Council and <http://www.guideline.gov/> by the US Department of Health and Human Services, Agency for Healthcare Research Quality.

of data consist of images; the second largest quantity of data found in medical records is the data gathered automatically from various monitoring devices (e.g., for heart rate, blood pressure, and breathing) as well as other care devices. The smallest quantity of data found in such records consists of typed data. Even this data size is obviously not small. In fact, during one single inpatient period in intensive care, the data typed as free-form text alone can translate up to 37,000 words or about 75 pages (Suominen & Salakoski 2010).

In comparison, the publicly available, uncurated information originating from Twitter for professionals and laypersons grows by 400 million tweets per day (Twitter Blog 2013). This corresponds to 20 million pages of new text every day. Reading them all would take over 60 years.

Health information written by laypersons to their peers on social media sites may be uncurated, yet it captures the valuable personal experiences of patients, patients' next-of-kin, and their caretakers with regard to health issues and access to health care. People use the Internet and social media, especially Twitter, increasingly to search for knowledge and uncover trends related to health and health care (Young & Bloor 2009). Nearly half of Europeans consider the Internet as an important source of health information (Kummervold et al. 2008). More than 80% of Australians use the Internet, and over 40% of Australian searches on the Internet are related to health and medical information (Experian Hitwise 2008; The World Bank 2013). An online survey about the use of health related social networking sites found that 85% of the survey participants were seeking information about their medical condition on line (Colineau & Paris 2010). Approximately three out of every four US adults use the Internet, and 80% of them search the Internet for health information in particular (Fox & Jones 2009; Fox 2011).

6.2.2 Information search on social media

Social media supports information search through *user-entered keywords*. With the 200 million active users of Twitter creating over 400 million tweets per day, these active users send out 1.6 billion queries on Twitter per day (Twitter Blog 2011, 2013). To support these information seekers, social media platforms have specified and implemented, for example, *hashtags* on Twitter (e.g., *#healthcare*) and the *<meta> field* in an Internet page to insert a short description of the page, keywords, and other metadata. However, their availability is conditional to users entering the appropriate keywords.

Since 2011, Twitter has also supported information search by creating *personalized search experiences* based on using information about the information seeker (e.g., geographic location, preferred language, and social relationship

between the seeker and author). This personalization is a timely topic, because the total personalized healthcare market in the US alone has been predicted to double from approximately 250 billion USD in 2009 to 500 billion USD in 2015 (Price Waterhouse Coopers 2010).

However, this current support for information search is not enough to leverage collective knowledge in social media. In other words, when considering social media, especially Twitter, for use in key decision making, we need to recognize three facts:

First, using Twitter content is comparable to “drinking from a firehose” (where an onslaught of information is coming at a person all at once, and they have to struggle to be able take it in all), or, alternatively, “searching for a needle in a haystack” (where a plethora of information is present but one has no clue where to find the kernel of information they seek). Twitter’s usage is growing exponentially over time (Sullivan 2010; Twitter Blog 2011, 2013). Here is how: In 2010, an average of 600 tweets were sent per second. By 2011, this had nearly quadrupled to over 2200 tweets per second. By 2013, the number of tweets per second had grown to 4600. When considering a particular search topic, this rate of tweets per second can vary substantially, and the reason for the wide range in the number of tweets on various subjects can be equally mystifying.

For example, the rate of tweets per second peaked at over 140,000 (i.e., about 7000 pages per second) in August 2013 in response to viewers’ discussions on a particular Japanese television show (Twitter Blog 2013). The explanation of this record rate of tweets was that viewers during the Nippon Television Network’s airing of *Laputa: Castle in the Sky* joined the animation film’s protagonists in casting a magic spell, known as “BALS.”⁷ Since viewers needed to participate in casting this spell in simultaneity with the film characters, so that they could assist the protagonists in closing down or destroying the city of Laputa, the number of tweets broke all records in Twitter history (Madoka 2013). In contrast, when looking at stories that are less theatrical or mystical, even though they may have an effect on large populations of community residents, the rate of tweets per second shows a vastly different picture.

When combining the number of tweets from the New Zealand earthquake and the Australian floods, the total number of tweets that were generated was 52,600 over 3 months (Bruns et al. 2012; Kreiner et al. 2013; Twitter Blog 2013). This number may sound small when compared with 140,000 tweets per second, but it still corresponds to over 2600 pages of text. As a further complexity, tweets are

⁷ Laputa: Castle in the Sky is a popular Japanese film that involves an animated cartoon, often with violent or sexually explicit content. In the language of Laputa, “BALS” means “close” as in “close down.”

authored by the 200 million active users who are registered with Twitter, and the Twitter service handles 1.6 billion search queries per day (Twitter Blog 2011, 2013). Consequently, such linguistically short-spoken and lexically extensive data needs to be assessed, related and visualized with respect to all different search topics, information needs, and query variants. In the presence of these requirements, keyword and/or hashtag searches on Twitter are insufficient, and often prone to returning false information. Our key observation is that search tools must be very effective at extracting meaningful content from the vast amounts of raw data.

Second, until recently, standard searches on Twitter have been limited to tweets that are not older than approximately a week (Sullivan 2010). The reason for this is that scalability must be the key performance measure in the evaluation of search engines for Twitter. Because earlier engines did not scale out (thus not enabling them to process vast amounts of raw data), they limited their data to recent tweets. However, with health issues such as infectious diseases, limiting the processing of tweets to a narrow window of time does not allow for the building of a knowledge base about important health trends. Thus, we cannot stress strongly enough how important it is for engines to scale out to the large numbers of tweets. This has motivated us and other R & D teams, as well as Twitter itself, to enhance the state of the art in the scalability of search engines. Consequently, in 2010–2013, significant progress has been made, enabling the processing of tweets over increasingly longer periods of time. Since 2011, the search engine of Twitter can search from the entire tweet history using the “divide and conquer method” (Twitter Blog 2011). Instead of analyzing the entire tweet history at once, the set of all tweets is divided to subsets based on the message posting-time. Then the engine progresses through these batches retrospectively from the most recent to the oldest tweets, accumulating new hits to the end of the result list every time the analysis of yet another subset becomes available. This method takes time, however, and the information seeker must thus wait a bit if there are a large amount of relevant tweets. In addition, the results are ordered strictly by time rather than by the relevance of such tweets. This can make it especially hard to find information right away that is directly relevant to the information seeker’s search query, since the user must first go through all the search results that are brought up because of their chronology. From what we describe here, it is understandable that a good deal of more work is needed to address this important search issue.

Third, tweets’ information quality and reliability needs to be monitored, and this text analysis needs to be combined, compared, and correlated with other sources of health information. The content of relevant tweets (assuming that the many irrelevant tweets have been removed) may be erroneous or even malicious. Some analysis approaches have already been shown to combat effectively the

so-called “fake” imagery in Twitter (e.g., during Hurricane Sandy false images of sharks swimming in suburban streets were created and tweeted), and general false statements such as those that occurred at the time of the Boston Marathon bombing (Gupta et al. 2013a, b). In addition, tweets provide extensive metadata that can inform automated analysis technologies (Lim et al. 2013). In conclusion, our key observation is that obtaining useful reliable health information from Twitter requires highly efficient search and processing mechanisms. While work has been done towards this goal, more research is required before Twitter can be used effectively to gather informative content.

6.3 Proposed solutions

Social media promises “collective knowledge” or “wisdom of crowds,” where many small observations may be aggregated into large, useful information. Search engines with increasing intelligence are needed to leverage this collective knowledge in social media. First, as mentioned in the prior section, tweets need to be assessed, related, curated, and visualized with respect to search topics, queries or emergent events (anomalies). Second, this text analysis needs to be combined, compared, and correlated with other sources of information.

The task of building an Internet search engine that retrieves documents whose content matches a search query is called *information retrieval* (Manning et al. 2008). Typically, the related solution techniques belong to the field of *machine learning* (Dua et al. 2014) that enables computers to learn to carry out specific tasks. Often this learning is the result of observing the connections between data and desired solutions. Also, sometimes techniques belonging to the field of *natural language processing* (NLP) (Nadkarni et al. 2011) are employed. NLP refers to automated techniques for understanding and generating a human language.

In the following three subsections, we describe some solutions that have been proposed to improve Twitter search engines, either by improved understanding of laypersons’ experiences or through contributions to information curation.

6.3.1 Tools for information retrieval on twitter

Twitter supports information search through user-entered keywords (i.e., hashtags) and personalization (i.e., use of geographic location, preferred language, social relationship between the seeker and author, and other information about the information seeker). However, this current support for information search is conditional to users entering the keywords and insufficient to leverage collective

knowledge in social media. The question is, then: How can we build a basic search engine and then increase its intelligence?

6.3.1.1 Basic recipe for building a search engine

A basic recipe for building a search engine is to first *create a repository of indexed documents to be searched for* and, second, *implement the search functionality*. This can be accomplished in many ways, but next we provide a basic recipe.

The repository of indexed documents can be an existing collection of electronic documents, or it can be created by using an *Internet crawler* (a.k.a. spider, ant, scutter, or indexer) that is initialized with a given Internet page and used to follow every link on that page respectively, and then store the new documents that are found (Kobayashi & Takeda 2000). The Twitter *Application Programmer's Interface* (API) provides a way to collect a repository of tweets.⁸ The collection can be initialized by user names; tweet *identification numbers* (id); hashtags or other *search terms* (a.k.a. query terms); geographic location; or time. The step of indexing the documents specifies an id for each document in the repository and stores the content that will be used by the search functionality. That is, the functionality can consider all document content, or it can be limited to, for example, the title, abstract, metadata, or the page's *uniform resource locator* (URL) in the Internet.

The search functionality implements a way for the seekers to enter queries, eliminate duplicate documents, assess the relevance between a given query and document, return relevant documents, and *rank* the returned documents in the order of most-to-least relevant. For example, the *Apache Lucene* is a piece of open-source software that developers can use for this search task.⁹ To make the functionalities more intelligent, they can also assess the trade-off between the content width and depth in returned documents. This trade-off means that all returned documents should *not* describe the same content, but rather supplement each other by providing different kinds of content. For example, the query of *obs* may refer to obesity, obsessiveness, obstetrician, and Osteoblasts, just to name a few; and *RR* to radiation reaction, recovery room, relative response, respiratory rate, and risk ratio, among others. Or more detailed content can be provided for that matter as in “an *obstetrician* is a medical professional” versus “a physician who provides care for women and their children during prenatal, childbirth, and post-natal periods.” Moreover, increasing intelligence can be used in visualizing the search outcomes in more interpretative ways than by just providing a ranked list of documents.

⁸ Twitter API (<https://dev.twitter.com/docs/api>)

⁹ Apache Lucene (<http://lucene.apache.org/>)

In order to improve the search results with Twitter data, it may be beneficial to use Twitter-specific tools to process the documents before using the search functionality. Example tools include:

1. *TwitterNLP* (Gimpel et al. 2011) for *tokenization* (i.e., identifying words in a stream of text and replacing the words with their base forms) and *part-of-speech tagging* (POS, i.e., tag every word in a stream of text with its grammatical category, such as noun, punctuation mark, or abbreviation);
2. *Twitter stopwords* (i.e., a list of common words not to be used in indexing and search); and
3. *Twitter-sentiment-analysis* for performing *sentiment analysis* (a.k.a. opinion mining, which identifies and extracts subjective information in documents such as when a tweet is describing a subjective vs. objective statement, or is expressing a positive, neutral, or negative opinion).¹⁰

6.3.1.2 Solutions

Su et al. (2011b) applied machine learning and NLP to searching information in Twitter in order to support consumers of health information in finding relevant concepts and in discovering new knowledge and trends. The search engine, called *TweetDetector*, is made more intelligent by adapting *query expansion* (QE) (i.e., adding search terms to a query by including synonymous concepts) and by *ranking* (i.e., returning the output list in the relevance order, as explained in the previous subsection) for the purpose of information retrieval.

First, the search engine performs *normalization* of tweets and the query that elicits the tweet responses. This aims to improve lexical coherence, expedite the search, and save memory on a computer. The normalization begins by removing stop words using the *SMART system*,¹¹ supplemented with www and other repeatedly used terms in social media. The normalization continues with *LingPipe*¹² in order to extract the tweet message for query expansion. Finally, words in the tweet and query are lemmatized using the *Porter Stemmer*.¹³

Second, the search engine performs query expansion on the normalized tweets and queries. This is accomplished by a series of steps: (1) asking an information seeker to enter an *original search query*; (2) *retrieving tweets* that include

¹⁰ TwitterNLP (<http://www.ark.cs.cmu.edu/TweetNLP/>), Twitter stopwords (<https://code.google.com/p/twitter-sentiment-analysis/source/browse/trunk/files/stopwords.txt?r#equal#51>), and Twitter-sentiment-analysis (<https://code.google.com/p/twitter-sentiment-analysis/>)

¹¹ SMART system (<ftp://ftp.cs.cornell.edu/pub/smart/>)

¹² LingPipe (<http://alias-i.com/lingpipe/>)

¹³ Porter Stemmer (<http://tartarus.org/martin/PorterStemmer/>)

at least one original search term; (3) *identifying from the retrieved tweets nine additional terms* to be included in the query; and (4) *repeating the search with this expanded query*. Let us now clarify the last two steps.

At the third step, the engine defines the additional terms by using so called *TFxIDF* and *Rocchio's QE with pseudo feedback* models, which are among the default choices in information retrieval (Manning et al. 2008). The *TFxIDF* model calculates a term weight v_t for each unique term t in the set of retrieved tweets, and the model of *Rocchio's QE with pseudo feedback* re-scales these weights. *TF* refers to the *frequency of term t* in the set of retrieved tweets and *IDF* to *inverse document frequency*, defined as the total number of retrieved tweets divided by the number of tweets containing the term t . In the *TFxIDF* model, these two numbers are scaled logarithmically and multiplied. These *TFxIDF* weights v_t are then re-scaled as defined by the model of *Rocchio's QE with pseudo feedback*:

$$w_t = \alpha v_t + \frac{\beta}{r}(t_1 + t_2 + t_3 + \dots + t_D)$$

where r refers to the number of relevant tweets and is set to 20; t_d records whether the term t is mentioned at least once in the tweet d ; D is the total number of tweets; and the default values of 1.00 and 0.75 are used for the parameters α and β , respectively. Using this re-scaling does *not* require feedback from the information seeker – with pseudo feedback, the tweets that are ranked in the top 20 by the original search are assumed to be relevant.

At the fourth step, the search engine returns the expanded query with the original query and the terms associated with the nine largest weights. The implementation of all these steps is based on the aforementioned Lucene search engine.

This simple engine performs significantly better than an even simpler engine that is otherwise the same but ignores the query expansion step. Intuitively, tweets containing more search terms (or expanded terms) are ranked higher. The retrieval performance of this simple approach, that is, the quality of the search results, varies from satisfactory to excellent, depending on the search terms. This performance evaluation was based on the set of a million tweets; namely, queries of *diet*, *weight*, and *weight loss*, and two information seekers. The comparison uses the *Precision at N ($P@N$)* measure (Manning et al. 2008), defined as the proportion of the relevant tweets that are retrieved out of the top N retrieved tweets. This measure takes values between zero (i.e., the worst performance) and one (i.e., the best performance). For N , the evaluation uses the values of 10, 20, 30, ... , 100.

Su et al. (2011a) extended the engine by considering other approaches for tweet processing (Fig. 6.2). The extension aimed to improve the quality of the search results by increasing the intelligence of the engine. The development includes *synonymy analysis*, *link expansion*, and *active learning*.

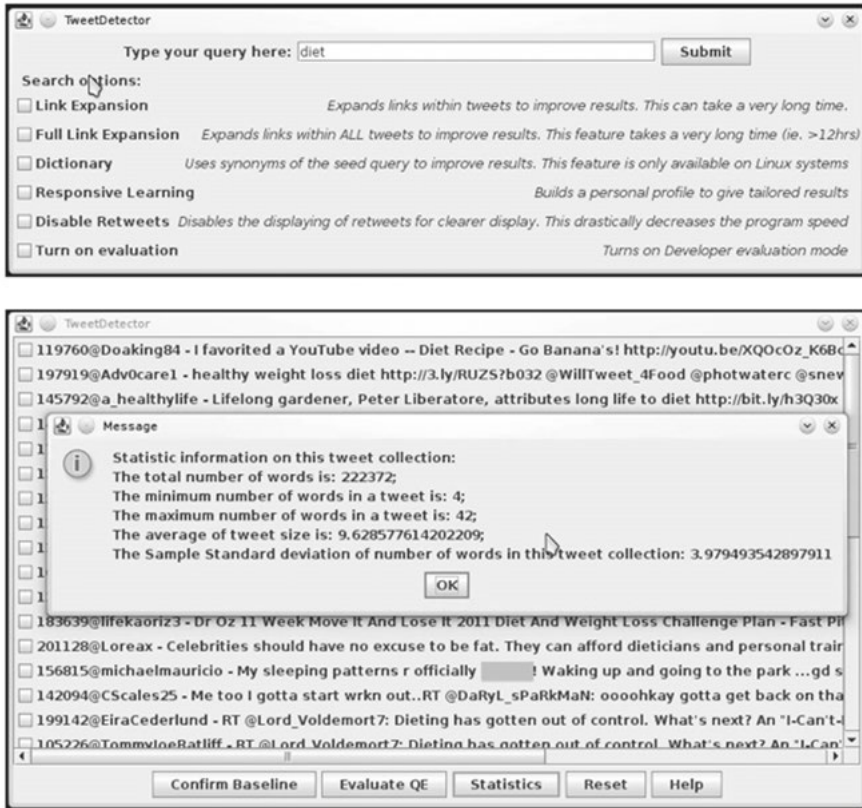


Fig. 6.2: Screen captures from using the extended TweetDetector engine for the query of diet.

The synonymy analysis component extends the aforementioned QE approach by appending the search query with more content. It is performed before applying other components. The component uses the *Infomap NLP Software* to identify terms used in tweets that could be synonymous to the search terms.¹⁴ Synonymous terms are those that have a synonymy score larger than a given threshold value (i.e., 0.7 in this case); the score values are between zero (i.e., very dissimilar terms) and one (i.e., very similar terms). If synonymous terms are found, the query is enriched by appending it with these synonyms.

The link expansion component enriches the tweet messages by appending them with more content. First, it uses the *URL class of java.net* to identify all

¹⁴ Infomap NLP Software (<http://infomap-nlp.sourceforge.net/>)

URLs mentioned in tweets.¹⁵ Second, it follows the URLs and analyzes both the <meta> field of the Internet page and the rest of the page. The parts considered in the analysis of the <meta> field include the keywords, description (i.e., an overview of the page content), and content language (e.g., *en* for pages in English). As discussed above, the availability of these parts is conditional to the page creators entering the content. The paragraph analysis uses the *CyberNeko HTML Parser* to parse the page in *HyperText Markup Language* (HTML) and access its information as a tree in *Extensible Markup Language* (XML).¹⁶ This enables efficient segmentation of the page into paragraphs. Based on applying rules¹⁷ to the paragraph text, the analysis determines whether a given paragraph contains *key content*. If the content language field indicates that the page is in English, the tweet message is enriched by appending text found under *keywords*, *description*, and *key content*.

The active learning component attempts to better personalize the search results by exploiting tweets that a given information seeker has previously marked as relevant. First, the component removes duplicate tweets (e.g., forwarded messages) and very similar tweets (e.g., messages that have been forwarded after a slight modification of the content). To define this similarity, the component analyzes the re-scaled weights w_t for two tweets, and, if their difference is small, the tweet with the smaller value is removed. If the values are equal, the second tweet is removed. Second, the information seeker is shown the resulting list of the top 20 retrieved tweets. Third, the seeker evaluates the relevance of these tweets to the query by marking the relevant tweets. Fourth, the term weights v_t are adjusted accordingly, and the rest of the search process is repeated. This results in revised search results (i.e., tweet listing).

The retrieval performance of the extended system has been compared against the original TweetDetector engine using the set of 300,000 tweets; queries of *diet*, *weight*, and *weight loss*; and four information seekers. The comparison uses the $P@100$ measure. In general, the extension seems to contribute to the performance (Fig. 6.3), but it takes time and requires feedback from the information seeker.

¹⁵ URL class of java.net (<http://docs.oracle.com/javase/6/docs/api/java/net/URL.html>)

¹⁶ CyberNeko HTML Parser (<http://sourceforge.net/projects/nekohtml/?source#equal#navbar>), HyperText Markup Language (see, e.g., the tutorial <http://htmldog.com/guides/html/>), and Extensible Markup Language (see, e.g., the tutorial <http://www.w3schools.com/xml/>)

¹⁷ For example, the rule of including at least w words and c commas is a heuristic to identify if the page creator has entered a paragraph that looks like a comma-separated list of words. Because this is likely to be a list of keywords, the text is added to the key content list.

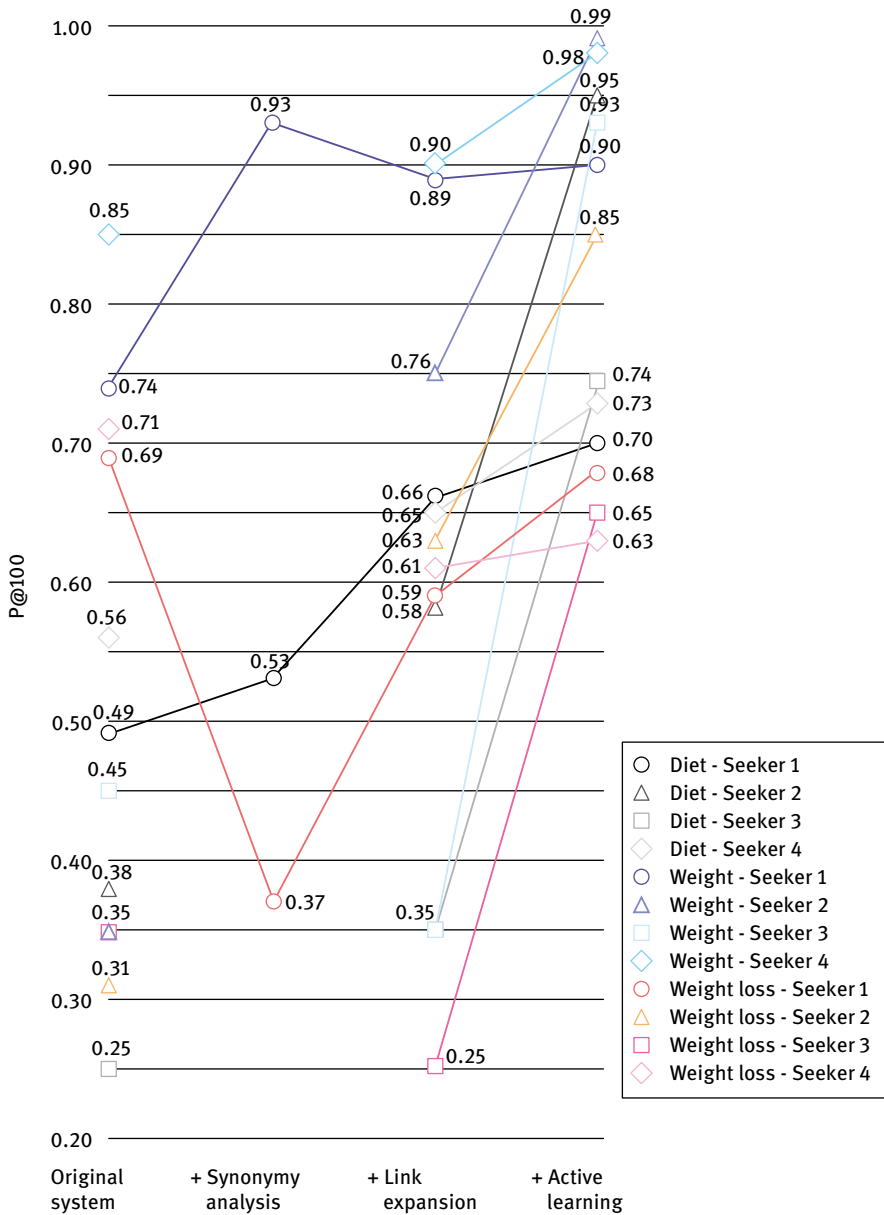


Fig. 6.3: Retrieval performance for the queries of diet, weight, and weight loss. The statistics to generate the graph originate from Su et al. (2011a).

6.3.1.3 Health concerns, availability of clean water and food, and other information for crisis management knowledge from twitter

Social media is a major source of crisis information in 2013. For example, in Australia, over 80% of people use the Internet, three out of every four people can access the Internet on a smartphone or other mobile device, and this accessibility has more than doubled from 2010 to 2013 (CampaignBrief 2013; The World Bank 2013). The potential of social media to help with crisis management has been already highlighted – see, e.g., Griffen et al. (2012). But is this a reliable source, and how can we curate this information?

6.4 Background

The role of Twitter changed from a personal network for messaging in real time to a global *crisis informatics solution* in January 2009. Crisis informatics is defined as the interconnectedness of people, organizations, information, and information and communication technology during a crisis situation (Hagar 2006). On January 16, 2009 at 7:33 AM, a tweet broke the news about a plane crash in New York 15 minutes before the media (The Telegraph 2009). Ever since, Twitter has gained increasing importance during natural disasters, epidemics, and civil unrest. Let us consider the following three examples over the past 6 years:

First, in California in October 2008, when wildfires destroyed 500,000 acres of land and 1500 homes, three out of every four people sought information on the Internet, and over a third of people created or shared information on Twitter or similar Internet solutions (Sutton et al. 2008).

Second, in Australia and New Zealand in January–March 2011, during the Queensland floods and the Christchurch earthquake, there was, likewise, a fair amount of tweeting (Bruns et al. 2012; Kreiner et al. 2013). The flooded area in Australian Queensland was 1,000,000 km², which is more than the combined area of states of Texas and New Mexico in the US or of the European countries of France and Germany. The New Zealand earthquake killed 185 people and costs of re-construction were estimated as USD 17–25 billion. Over 9400 people created over 52,600 unique tweets related to the earthquake, with nearly 860,000 words (approximately 75,800 unique words) in total, and the floods drew approximately 1100 tweets every hour.

Third, the AH1N1 (a.k.a. swine flu), the Japanese crisis, and the Libyan civil war took a major toll on human life. In the aftermath of the 2011 AH1N1 flu outbreaks in South and Central America as well as in the Caribbean and Mexico, the *World Health Organization's Pan American regional office* issued a pandemic preparedness alert because of the vast number of people affected by the influenza outbreak.¹⁸ The “Japanese crisis” refers to a catastrophic chain of events back in March of 2011, when an earthquake of a magnitude of 9.0, considered the fifth most

powerful earthquake in the world, killed nearly 16,000 people in the surrounding Tokyo region, triggering a tsunami, which claimed another 3500 lives. A nuclear disaster ensued when the flooding from the tsunami paralyzed the backup generators that were supposed to prevent the fuel rods from overheating. This caused a nuclear incident that was considered among the 10 worst nuclear accidents in history, similar to Chernobyl in 1986 and Three Mile Island in 1979 (Cohen 2011).¹⁹ The Libyan civil war, in turn, killed at least 30,000 people, injuring another 50,000 in their six-month long civil war (Laub 2011). Twitter no doubt became the pulse of these events, with the AH1N1 flu being the lead topic on Twitter, followed by the Japanese crisis and the civil war in Libya (Twitter Blog 2011).

6.5 Some solutions

Tweets hold the potential to be useful in crisis management, including the management of health related crisis (e.g., the AH1N1 crisis mentioned above). But to fulfill that potential they need intelligent search and summarization systems to assure their accessibility and reliability. Both NICTA (National ICT Australia) and CSIRO (Commonwealth Scientific and Industrial Research Organization) have developed such systems. We describe below the two NICTA systems in some detail, while we present only a summary of ESA, a system developed by CSIRO, and refer the reader to more detailed descriptions in Yin et al. (2012); Cameron et al. (2012); Power et al. (2013); and Karimi et al. (2013).

Crisis managers, media and other people interested in summarizing Twitter in real time can use *NICTA EventWatch* to monitor the message topics.²⁰ The summary aggregates messages into meaningful clusters as they emerge over time. That is, this system is designed for monitoring Twitter in real time. Details of the *clustering* method used in EventWatch can be found in Can (1993), and more recent methods that could be useful are described in Mehrotra et al. (2013). Also the live demonstration of EventWatch is available in the Internet (Figs. 6.4 and 6.5).²¹ The demonstration uses the query of *#auspol lang:en* corresponding to tweets containing the hashtag *#auspol*, for Australian political discussion, in English.

¹⁸ World Health Organization for the Pan America (<http://www.paho.org/>) and the AH1N1 alert (http://new.paho.org/hq/index.php?option#equal#com_content&task#equal#view&id#equal#5291&Itemid#equal#1091&lang#equal#en)

¹⁹ See, e.g., http://en.wikipedia.org/wiki/Fukushima_Daiichi_nuclear_disaster and references therein.

²⁰ NICTA EventWatch (http://www.nicta.com.au/business/broadband_and_the_digital_economy/projects/eventwatch)

²¹ EventWatch demonstration (<http://eventwatch.research.nicta.com.au/demo/#>)



Fig. 6.4: EventWatch area graphs without (i.e., top) and with (i.e., bottom) the sentiment analyzer.

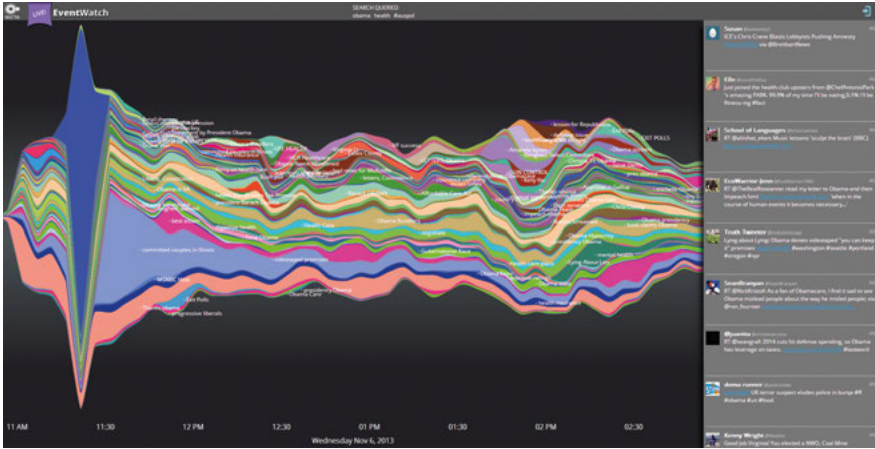


Fig. 6.5: Stream graph illustrating the topic distribution in time on EventWatch.

Let us illustrate a use case of EventWatch as follows: Consider an information seeker with domain knowledge who requires immediate information on emerging topics of interest. For example, during a crisis, an information seeker could have the following information need/query: *“what are the key requests for resources”*; or during a technology conference, the topic of interest might be: *“what are the emerging questions or focus points?”* In both cases, the seeker needs to quickly sift through social media streams (e.g., Twitter) and develop a summary of activity (i.e., what are the key topics) along with an action plan (i.e., how to respond) and then to support other activity – typically by providing situational advice. The technology is expected to operate in real-time, providing a visual summary of topics as an alert to the already busy seeker. EventWatch is not an automated decision support tool, but rather a technology assistant that allows the seeker to focus on key areas of interest: although machines cannot replace people in understanding the text, they can reduce the stream to manageable topic areas for further investigation.

EventWatch operates on a query-based analysis of Twitter: the seeker defines a query (e.g., a hashtag, keyword, boolean query, or a combination of these) which is then passed to Twitter. This user-defined query selects the tweets from the Twitter stream, which provides a near-real-time stream of matching tweets. The use of queries is due to the limitations of the current public Twitter API, which only returns 150 tweets per hour (around 1%) from the complete Twitter stream. A query-based system improves the chances that the tweets found will be useable, and it is recommended by Twitter to avoid applications being “rate limited” (blocked).

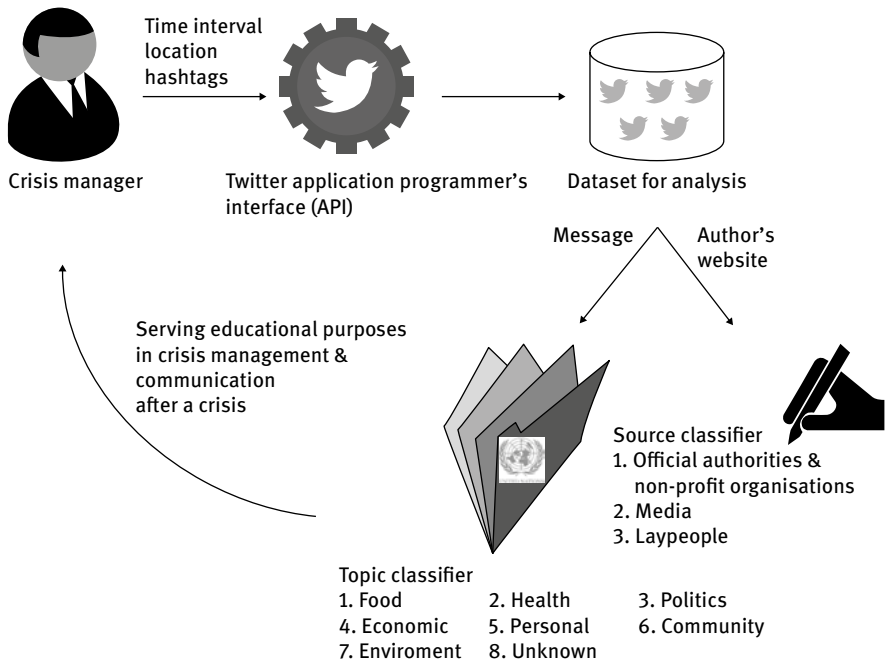


Fig. 6.6: Workflow for using the NICTA automated classifier for crisis managers.

The text of the tweets in the stream is then analyzed to extract keyphrases. Keyphrases are formed from two or more *unigrams* (i.e., an *n*-gram is a text sequence of *n* words, and thereby, a unigram is a one-word sequence such as *opening* or *keynote*) that frequently co-occur in a text corpus (e.g., *opening keynote*, *live demo*, or *stunning animations*). These keyphrases are generally more useful than single words because they are more specific. The keyphrases define topics that may be of interest to the information seeker. Given topics can grow or shrink depending upon the scale of activity in the Twitter stream.

Within the topics, *named entities of people, places, and organizations* can be extracted, and sentiment (i.e., if a tweet is positive or negative) can be found. The topics themselves are linked to the live Twitter stream, to allow the seekers to draw their own conclusions about the quality of the analysis. The technology allows the seeker to view emerging topics within a specific area of interest. If the seeker clicks on a given topic, they can perform a forensic analysis of the tweets related to the topic.

The *NICTA automated classifier for crisis managers* analyzes tweets retrospectively in order to classify them with respect to topics and information sources (Kreiner et al. 2013; Fig. 6.6). Results serve educational purposes in crisis

management and communication after a crisis. For example, the distribution of messages across the topics (sources) can demonstrate the nature (reliability and maturity) of the crisis. When applied periodically over time, analyses could reveal early indicators for specific developments, trends, and effects of crisis communication during a crisis.

The *topic classifier* uses seven threat categories by the United Nations (i.e., *Food, Health, Politics, Economic, Personal, Community, and Environment*), supplemented with the categories for *Other* potentially relevant information to crisis management and *Irrelevant*. The categories for information relevant to crisis management are defined as follows: If the tweet is relevant to food quality or supplies (e.g., availability of clean water and food), assign it to the category of Food. Otherwise, consider the other categories in the order of Health, Politics, Community, Personal, Economic, Environment, and Other. If none of these categories is relevant, assign the tweet to Irrelevant.

First, the topic classifier uses the aforementioned Twitter POS Tagger for lemmatization and part of speech tagging. Second, it replaces more specific terms with more general terms (e.g., food, natural phenomena, and possession) and expands shorthand using the *WordNet on NLTK*,²² supplemented with references to images, numbers, money, geographic locations, Twitter users, hashtags, Internet addresses, and shorthand. Third, it extracts physiological features using the *Regressive Imagery Dictionary on NodeBox*.²³ Fourth, it applies the optimal subset of 72 binary features to tweet classification using *Naïve Bayes on Orange*.²⁴ The classification performance is good in the major categories of Economic and Community, but more modest in the minor categories, where not enough data were available for a proper initialization. The evaluation is based on the set of over 100,000 tweets, of which 1000 are evaluated by two information seekers. The measures of the number of *true positives* (i.e., tweets for which both the seeker and classifier assign the same category *c*), *false positives* (i.e., tweets for which the seeker does not assign the category *c* but the classifier does), *true negatives* (i.e., tweets for which both the seeker and classifier do not assign the category *c*), and *false negatives* (i.e., tweets for which the seeker assigns the category *c* but the classifier does not) for each category *c* are applied.

The *source classifier* detects whether the information source is: (1) an *official authority* (e.g., emergency agencies, ministries, police, fire services, military

22 WordNet on NLTK (<http://nltk.org/>)

23 Regressive Imagery Dictionary on NodeBox (<http://nodebox.net/code/index.php/Linguistics>)

24 Naïve Bayes on Orange (<http://orange.biolab.si/>)

sources, the Red Cross or the Green Cross,²⁵ or any other non-profit organization); (2) the *media* (e.g., online magazines and television/radio stations); or (3) a *lay-person*. It is rule based; that is, the analysis focuses on presence or absence of 123 class-specific top terms. The classification performance is excellent in all categories. The evaluation is based on the set of over 100,000 tweets, of which 600 are evaluated by one information seeker. Again, the measures of the number of true positives, false positives, true negatives, and false negatives for each category *c* are applied.

In contrast to these NICTA technologies, the *CSIRO ESA System* (Cameron et al. 2012; Yin et al. 2012; Karimi et al. 2013; Power et al. 2013) is not query-based. It constantly monitors the Twitter stream and performs a real-time analysis (1) to look for an anomaly, which could reflect a specific event, a crisis or an emergency that needs attention, or (2) to identify that a specific event, such as a fire, has happened and requires a response. An anomaly is defined as a variance in the type of messages or the information being posted. When an event occurs and many people post about it, the words associated with the event will be mentioned more frequently than they normally would. For example, an epidemic might result in many people posting about the disease. ESA uses a *burst detection algorithm* that analyzes incoming data in real-time. It can then produce a variety of visualizations of the discovered bursts (see Figs. 6.7–6.9) to enable someone to be immediately alerted to it. ESA clusters messages according to their topics so as to provide the person monitoring social media with a useful overview. ESA is also able to identify, in real-time, tweets that indicate specific crisis events, such as fires, through both a set of heuristics and a classifier (Karimi et al. 2013; Power et al. 2013). When encountering tweets that mentions such an event, ESA further processes them to determine if they correspond to an actual crisis event or not,



Fig. 6.7: Burst visualization in ESA for a small earthquake in Melbourne in early 2014. The color and size of the word indicates the size of the burst. (Screenshot provided by Bella Robinson and Mark Cameron.)

²⁵ An environmental organization building upon the 1992 Earth Summit in Rio de Janeiro with chapters in over 30 countries.



Fig. 6.8: ESA: Heatmap showing the locations of tweets contributing to a red “christchurch” alert soon after an earthquake was felt there on 18 Nov 2013. (Screenshot provided by Bella Robinson and Mark Cameron.)

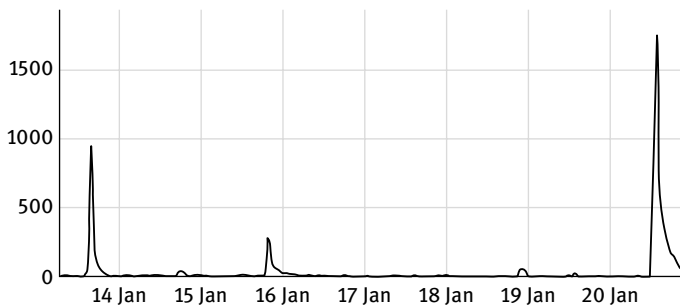


Fig. 6.9: ESA: Hourly volume of “earthquake” tweets early 2014, displayed on a timeline. The first spike corresponds to the earthquake in Melbourne as shown in Fig. 6.7, and the big spike on the right corresponds to an earthquake in Wellington. (Screenshot provided by Bella Robinson and Mark Cameron.)

thus helping someone with the problem of finding “a needle in a haystack,” and generating an alert (e.g., an email message with the pertinent information) to the persons monitoring for crisis events. Finally, ESA includes an ability to analyze the tweets retrospectively, for example, to understand trends. ESA is currently being used by a number of organizations, including crisis management agencies.

6.6 Tools for combining, comparing, and correlating tweets with other sources of health information

As highlighted above, in order to curate tweets' information quality and reliability, their contents need to be combined, compared, and correlated with other sources of health information. But what kinds of methods and tools can we use to accomplish this? Next, we present a summary of *Vizie*, a CSIRO system that achieves this and refers the reader to more detailed descriptions (Paris & Wan 2011, 2014).

The *Vizie* social media monitoring tool assists people to be able to monitor information-sharing on *all* social media platforms (i.e., not restricted to Twitter). This can be for a variety of domains, including health related information, and not necessarily limited to crisis management. In short, health information on social media is not restricted to Twitter or crisis management either. Health information can be shared on other social networking platforms, such as Facebook or MySpace, or specific online communities (e.g., PatientsLikeMe) and forums (e.g., *eHealth Forum* and *Consumers Health Forum of Australia*),²⁶ among others. Social media can be monitored for a number of health related issues, such as understanding how people talk about their physical or mental health, the medications they are taking, the treatments they are undergoing, or their diet. Social media can also be used to gauge people's reactions to the dissemination of official health related reports. The ability to look across various social media channels provides a more holistic perspective of the conversations that are happening on social media. It also enables the tool to compare the information gathered from tweets with other sources of information, thus addressing the information curation issue.

Vizie thus differs from TweetDetector, EventWatch and ESA in two ways: it collects data from all social media platforms, and it is not limited to crisis management. Like TweetDetector, it is query-based – that is, the information seeker needs to provide *Vizie* with one or more queries, and *Vizie* will search all incoming social media data for posts related to those queries. To monitor a wide variety of social media platforms, the *Vizie* system uses a *single federated search interface* that allows information seekers to understand why information is collected and where (e.g., which social media platform) the content came from. This source identification can be seen as serving information curation purposes in the same way as the aforementioned source classifier. It could be combined with a more general mechanism to classify sources into categories which would include, for example, healthcare organizations, media, and laypersons. Figure 6.10 shows a conceptual representation of the system.

Having collected data, *Vizie* employs NLP techniques to provide a unified analysis of the aggregated data. It uses *keyword and keyphrase detection*, with

²⁶ eHealth Forum (http://ehealthforum.com/health/health_forums.html), Consumers Health Forum of Australia (<http://ourhealth.org.au/>)

mechanisms based on Kupiec et al. (1995), *clustering algorithms*, based on keywords or Latent Dirichlet Allocation (LDA) (Blei et al. 2003), *discussion summarization* (Wan & McKeown 2004), and *extractive summarization* (Radev et al. 2003). These analysis are coupled with interactive visualizations and interfaces (see Figs. 6.11 and 6.12). Together, they support data exploration and help media

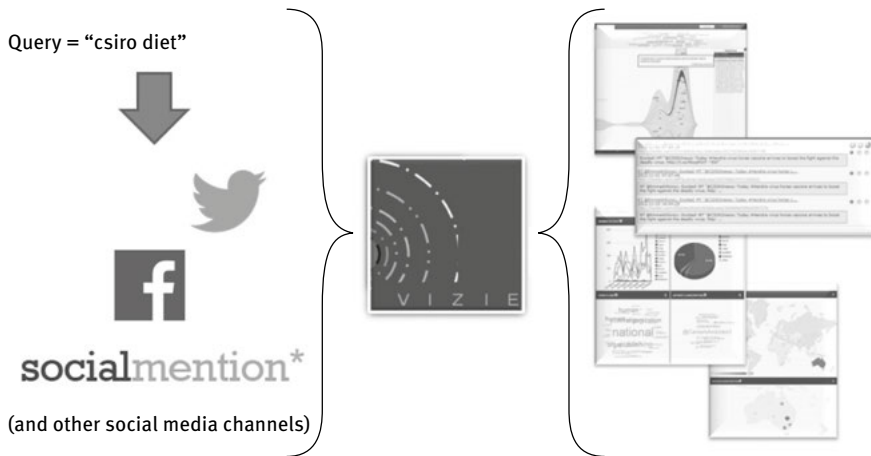


Fig. 6.10: Vizie: collecting and analyzing social media.

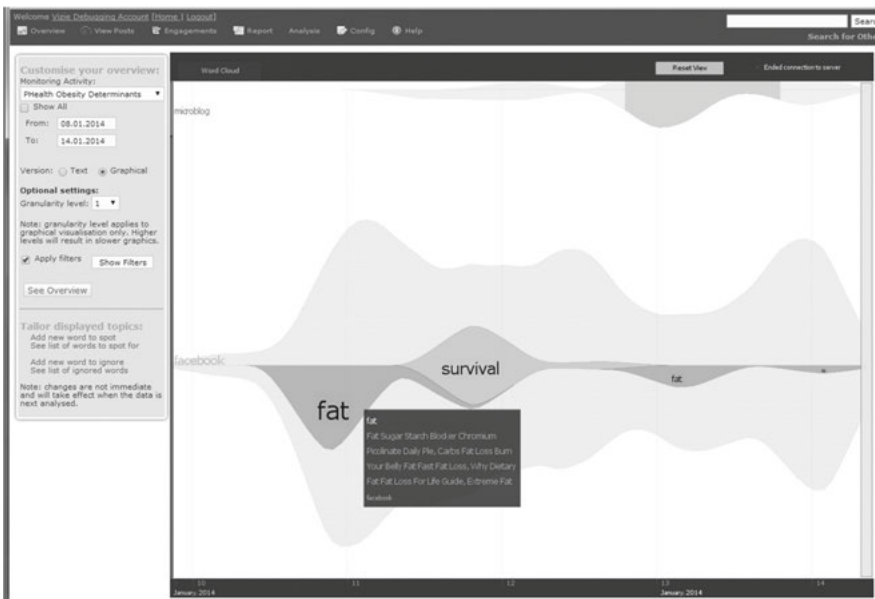


Fig. 6.11: VIZIE: a visual overview of social media about diet (zoomed on the Facebook data)

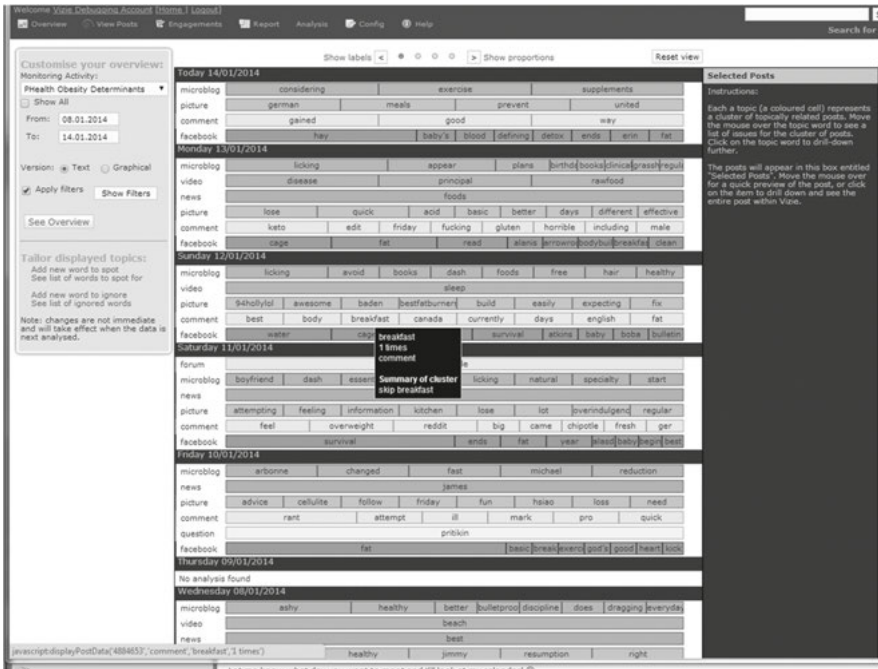


Fig. 6.12: VIZIE: a textual overview of social media about diet for a week.

monitors find commonly discussed issues across different types of social media. They also enable the cross validation of the information extracted on various channels, helping with information curation.

In addition, the Vizie prototype supports the seeker in deciding whether the content of a post is such that an engagement is required. It does this by assisting with relevance judgments, providing the context that triggers online discussions. This is illustrated in Fig. 6.13, where summary information from the linked document (accessible through the URL) is presented to the user as context for the tweet. Finally, Vizie also includes facilities for search (within the collected data), record keeping (archiving) and reporting.

Like ESA, the tool is being used by a number of organizations. Figure 6.14 shows the number of loggins in Vizie over the past 18 months (Wan & Paris 2014). The top blue line indicates the aggregate logins counts, while all the other lines are the counts for the individual organizations. One of the Vizie customers is an organization with a mandate and commitment to speaking to the broader Austrian community on issues of mental health and suicide prevention. The organization produces a report and is interested in finding out how people react to it, or more generally, how mental health issues are discussed in various communities.



Fig. 6.13: A summary of the news article, linked from the shortened URL, is presented to the information seeker as context to the Twitter discussion – (reprinted from Wan & Paris 2014).

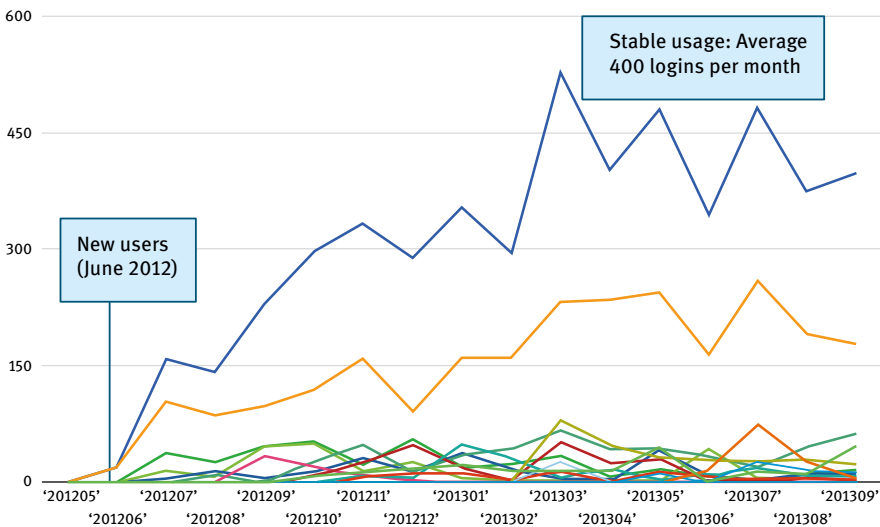


Fig. 6.14: Vizie: number of logins over 18 months (Wan & Paris 2014).

They have used the Vizie tool to obtain unique insights into what topics are being discussed and are of interest to a diverse range of groups.

6.7 Discussion

In this chapter, we have illustrated social media as a source of health information and positioned it with other sources. The use of Facebook, PatientsLikeMe, Second Life, Twitter, Wikipedia, YouTube, and other social media over the Internet is widespread among healthcare professionals and laypersons. For example, in 2013, PatientsLikeMe covers 2000 health/medical conditions for its 220,000 users to study, Twitter publishes 400 million messages per day for its 200 million active users, and these tweeters also enter 1.6 billion search queries per day. More generally, if considering the entire Internet, two-thirds of the US adults go on line for health information, almost half of the Australian searches relate to health, and nearly half of the Europeans consider the Web as an important source of health information. In addition to this aspect of offering a vital source of information for learning health knowledge, social media can enable people to interact and support each other emotionally during an illness. Other benefits include helping the society to realize the collective goal of improving healthcare outcomes and policies.

However, these new technologies have limitations too which we have addressed in this chapter by presenting the total of five systems that we have developed (three developed by NICTA, two by CSIRO) to help obtain useful information from social media. The main limitation of social media is its current inability to access and curate relevant information in the ever-increasing gamut of messages. Our systems improve search, summarization, visualization, and curation capabilities, both in real time and retrospectively, with the first four systems concentrating on tweets and the last one monitoring a wide variety of social media platforms. In essence, we have provided a basic recipe for building a search engine for social media and then made it increasingly more intelligent through smarter processing and personalization of the queries, tweets, and search results. We have also addressed the summarization aspect by visualizing topical clusters in messages and further classifying the retrieved tweets into topical categories that serve professionals in their work. Our solutions for information curation are to analyze the information sources and compare tweets against other sources of information.

6.8 Related solutions

We now mention some other solutions to the problem of mining social media for information.

6.8.1 Maps applications for disease monitoring

Mapping disease outbreaks across the world is a common use case in social media applications. For example, *Google Flu Trends* and *Google Dengue Trends* track seasonal outbreaks in the world, generating maps of disease activity on a five-point scale (from intense to minimal) by analyzing queries of the *Google Search* engine in the Internet (Ginsberg et al. 2008).²⁷ The analysis uses keyphrase matching to filter the relevant queries and the geographic location of the information seeker to map them.

HealthMap by the Boston Children’s Hospital, Boston, MA, USA, generates maps of disease outbreaks and warnings as well as surveys emerging threats to public health in real time based on the *EuroSurveillance*, *ProMEDMail*, *World Health Organization*, and eight other publicly available sources in the Internet.²⁸ It monitors a broad range of infectious diseases and, similarly to the Google solutions, the results are public on the Internet. The HealthMap page is also supplemented by the mobile application of *Outbreaks Near Me*.²⁹

Sickweather, in turn, considers messages and user profiles from Twitter and Facebook in order to track and visualize the prevalence of different illnesses, symptoms, and medical in a given area.³⁰ The connection with the user profiles enables finding sick friends and supporting them, for example, by sending a “get well soon” message.

6.8.2 Maps applications in crisis situations

Maps applications of this kind are also used in crisis situations. For example, in Kenya after the post-election fallout in 2008, 45,000 people used the *Ushahidi* maps for incidents of violence and peace efforts.³¹ People reported the incidents on the Internet or using their mobile phones.

The application was used again in January 2010 for an Haitian earthquake (Heinzelman & Waters 2010). This time, people could send information to the

²⁷ Google Flu Trends (<http://www.google.org/flutrends/>), Google Dengue Trends (<http://www.google.org/denguetrends/>), and Google Search (<https://www.google.com.au/>)

²⁸ HealthMap (<http://healthmap.org/>), EuroSurveillance (<http://www.eurosurveillance.org/>), ProMEDMail (<http://www.promedmail.org/>), and World Health Organization (<http://www.who.int/en/>)

²⁹ Outbreaks Near Me (<http://healthmap.org/outbreaksnearme/>)

³⁰ Sickweather (<http://www.sickweather.com>)

³¹ Ushahidi (<http://ushahidi.com/about-us>)

system as text messages, which volunteers supplemented with tweets and other social media messages from the Internet. The resulting public maps visualized medical emergencies; trapped people; and needs for shelter, water, food, and other necessities in real time. The United Nations and other authorities used them to conduct direct-assistance tasks such as resource allocation.

These maps applications can be extended to address more advanced modeling of topics. The extensions have been shown to result in a capability to distinguish distinct crisis situations of similar nature in nearby locations (Sumatran earthquake vs. Samoan tsunami), analyze the influenza AH1N1 evolution, as well as, more generally, explore prominent topics (Kireyev et al. 2009; Signorini et al. 2011). Predictive capabilities cover the detection of earthquakes, among other emerging events (Cataldi et al. 2010; Sakaki et al. 2010).

6.8.3 Extraction systems to monitor relationships between drugs and adverse events

Another example use of social media in health care is monitoring *adverse drug events* (ADEs), referring to injury or harm associated with the use of a given medicine. In terms of their prevalence, ADEs account for approximately 5% of Australian and US hospital admissions, and, among those admissions that result in an inpatient period in the USA, serious ADEs have been reported to occur in nearly 2% of the cases (Bates et al. 1995, 1997; Lazarou et al. 1998; DUSC 2000; Moore et al. 2007). Regardless of all efforts to decrease these percentages and the fact that from one- to two-thirds of all ADEs could be avoided by more careful prescribing and monitoring (Bates et al. 2003), the rates are unfortunately increasing four times faster than the total number of drug prescriptions to outpatients (Moore et al. 2007). If considering the severity of ADEs, in turn, they are within the five most common causes of death in Australian and US hospitals (DUSC 2000; Giacomini et al. 2007).

Next, we summarize two systems that extract relationships between drugs and adverse events from information posted on social media. Both systems are developed to study ADEs as reported by the patients themselves.

The first extraction system uses the comment text, disease name, drug name, and user id related to 3600 user comments for four drugs from the *Daily Strength* page on the Internet (Leaman et al. 2010).³² The processing studies the 3787 concepts of the *Coding Symbols for a Thesaurus of Adverse Reaction Terms* (COSTART),

³² Daily Strength (<http://www.dailystrength.org/>)

developed by the US Food and Drug Administration for post-marketing surveillance of ADEs; 888 drugs linked with 1450 ADE terms from the *Side Effect Resource* (SIDER); and associations between 10,192 drugs and 3279 ADEs originating from the *Canada Vigilance Adverse Reaction Database* or *Canada MedEffect2* resource for information drugs and health products.³³ It begins by coding colloquial phrases and grouping similar or synonymous meanings together manually. This is continued by automated merging of all concepts that contain a term in common into a single unified concept (SUC), dividing comments into sentences, tokenizing the sentences, associating the tokens with their POS tags using *Hepple* on the *Generic Architecture for Text Engineering* (GATE), removing the stopwords, and stemming the remaining tokens using the *Snowball* implementation of the Porter Stemmer.³⁴ Then, the processing scores the similarity between the comments and SUCs. The scoring covers each window of five tokens in the comment and each token in the SUC text. Finally, the scores are used to determine whether a specific concept is present in a given comment as follows: (1) the scores corresponding to this comment and the individual tokens in the SUC text are summed up; (2) the result is divided by the number of these scores; and (3) if this normalized score is greater than a configurable threshold, the respective concept is considered to be present in the comment. The concept extraction performance on two drugs is good: 78.3% of the retrieved comments are relevant to a given drug (i.e., *precision* of 0.783) and 69.9% of the comments that are relevant to a given drug are retrieved (i.e., *recall* of 0.699).

The second extraction system aims to identify ADEs from patient-provided drug reviews on health-related pages on the Internet with a focus on cholesterol-lowering drugs (Liu et al. 2011). It extracts side effect expressions followed by the construction of a side effect ontology using the total of 8515 messages that included drug reviews or use experiences from three drug discussion forums in the Internet: *askpatient.com*, *medications.com*, and *WebDB.com*.³⁵ The phrases of the dataset were manually classified into a hierarchical ontology. As a result, the 2314 side effect phrases were categorized into 307 synonym groups, which were furthered

³³ Coding Symbols for a Thesaurus of Adverse Reaction Terms (<http://hedwig.mgh.harvard.edu/biostatistics/sites/default/files/public/costart.html>), Side Effect Resource (<http://side-effects.embl.de/>), Canada Vigilance Adverse Reaction Database (<http://www.hc-sc.gc.ca/dhp-mps/medeff/databasdon/index-eng.php>), and Canada MedEffect2 (<http://www.hc-sc.gc.ca/dhp-mps/medeff/index-eng.php>)

³⁴ Hepple (<http://gate.ac.uk/gate/doc/plugins.html>), Generic Architecture for Text Engineering (<http://gate.ac.uk/>), and Snowball (<http://snowball.tartarus.org/>)

³⁵ *askpatient.com* (<http://askpatient.com/>), *medications.com* (<http://www.medications.com/>), and *WebDB.com* (<http://webdb.com/>)

grouped into 30 classes. Then, the system was used to extract 7500 shorter text snippets that included the side effect expressions from these phrases. After this, it removed from the snippet collection the 377 stopwords and expressions that did not occur more than five times in the dataset as well as ignored the word order (e.g., the snippets $word_1 word_2$ and $word_2 word_1$ are equivalent). Finally, the statistical analysis of the remaining 2314 unique snippets gave evidence for a significant correlation between the three drugs and a wide range of disorders and conditions, including amyotrophic lateral sclerosis, arthritis, diabetes, heart failure, memory loss, neuropathy, Parkinson's disease, and rhabdomyolysis.

6.8.4 An early warning systems to discover unrecognized adverse drug events

As opposed to extracting recorded ADEs from social media to derive statistics on incidence rates, the generation of early warnings has also been studied. The goal is to discover unrecognized ADEs faster and faster in order to warn healthcare professionals and drug consumers. Next, we describe four systems of this kind.

The extraction systems developed by Benton et al. (2011) and Brian et al. (2012) focus on drugs used to treat cancer. The former system uses the Porter Stemmer; *Cerner Multum's Drug Lexicon*; *Consumer Health Vocabulary*; hand-compiled vocabularies for dietary supplements, pharmaceuticals, and ADEs; frequency counts; and co-occurrence analyses.³⁶ Over 20% of the identified ADEs for four most common breast-cancer drugs from 1.1 million anonymized messages posted to eleven breast cancer message boards were new discoveries, that is, not documented on the drug labels. The latter system combines state-of-the-art classification techniques with synonym, negation, semantics analyses; word frequencies; and numbers of hashtags, reply tags, URLs, pronouns, and drug-name mentions. Its evaluation on two billion tweets sent May 2009 to October 2010 gives evidence of its good ability to identify ADEs correctly; over 70% of the detected ADEs were correct.

The demonstration software by Wu and Stanhope (2012) applies to a substantially wider range of medicine (not just cancer drugs). Its goal is also to generate early warnings. The analysis begins by identifying the strength of relatedness between two side effects with each drug. This uses *mutual information* to measure the amount of information one side effect carries about another (Manning &

³⁶ Cerner Multum's Drug Lexicon (<http://www.multum.com/lexicon.html>) and Consumer Health Vocabulary (<http://consumerhealthvocab.org/>)

Schütze 1999). It takes values between zero and one – the larger the values, the stronger the relatedness of the side effects. For example, the mutual information of zero means that the first side effect reveals no information about the other and vice versa. In contrast, the mutual information of one equals to the first side effect carrying all information needed to determine if the other effect also occurs. The analysis is finished by using the values of mutual information in a *hierarchical clustering* process (Manning & Schütze 1999). This results in a tree structure, where the leaves are single side effects and each node of the tree represents the group that contains all the side effects of its descendants. Similarly to family trees, where immediate descendants correspond to a person's offsprings and more distant relatives branch farther away, the tree of side effects visualizes the groups and relations between the effects. As examples of closely related side effects, the authors mention *Cerebral infarct* (i.e., a stroke resulting from issues in blood circulation to the brain) and *Status epileptics* (i.e., a life-threatening condition, where the brain is in a state of continuous or repetitious seizures), as well as *Allergic reaction* and *Tongue pain*. When using the aforementioned SIDER resource and Google discussions together with the medical Web pages of *MedlinePlus*, *Drugs.com*, and *DailyMed*, the system generated the average of 66 side effects per drug for 15,848 unique drugs.³⁷

Another warning system applicable to a wide range of medicine is introduced by Chee et al. (2011). Its goal is to generate a watch list of drugs for further monitoring of their safety by authorities such as the US Food and Drug Administration. The authors assembled a hybrid of Naïve Bayes and other state-of-the-art classification techniques to rank the drugs using specialized lexicons and word frequencies to define the features. The ranking applies a scoring measure defined as:

$$\frac{c f_+^2}{n}$$

where c refers to the number of classifier types in the hybrid; f_+ to the number of false positives, that is, cases in the evaluation dataset where the hybrid suggests incorrectly that a drug should be added to the watch list; and n to the size of this dataset. Multiplication by c penalizes the computational cost involved in including too many classifier types in the hybrid. The square brings the ability to differentiate the ratios of one false positive for a dataset of size two (i.e., 0.5 reflecting that classification errors are not very severe, because only a small dataset

³⁷ Google discussions (<https://groups.google.com/>), MedlinePlus (<http://www.nlm.nih.gov/medlineplus/>), Drugs.com (<http://www.drugs.com/>), and DailyMed (<http://dailymed.nlm.nih.gov/>)

was used in evaluation) from 100 false positives for a dataset of size 200 (i.e., 50, which indicates that the errors are a hundred times more severe because, when as many as 200 cases are considered, half of the predictions are false positives). The system was evaluated on a set of 12 million messages from 7290 public *Health & Wellness Yahoo! Groups*, including references to four drugs withdrawn from the market.³⁸ The evaluation tested the hypothesis of the hybrid being able to detect the withdrawn drugs even if they are intentionally labeled with the incorrect class (i.e., not to be added to the watch list); evidence for this hypothesis would be reflected in large scores, or equivalently, the hybrid generating many false positives. Indeed, the scores for three out of the four withdrawn drugs were 10.89, 10.89, and 10.24, corresponding to the ranks 4–6 in watch list. However, the fourth withdrawn drug had a low rank of 107 (score 0.04).

6.9 Methods for information curation

In the systems we presented, we have addressed the current inability of social media (in particular tweets) to curate information through a variety of means. This includes the automation of the classification of the information sources; identification of topics (including common topics) in the posts; detection of specific events; as well as combination, comparison, and correlation of tweets with other sources of information. To illustrate how crucial these capabilities are to filter out from search results all but topically relevant and reliable messages, let us return to the case of the Hurricane Sandy. Even if a given search engine was able to find the relevant tweets from the overflow of 340 million tweets a day,³⁹ false rumors and otherwise incorrect information in 91% of them is clearly a concern.

Others have attempted the automated analysis of the proportion of questions in tweet message chains (Mendoza et al. 2010). The data used in their study is a subset of the over 4.5 million messages on social media that are relevant to the Chilean earthquake in February 2010, killing 723 people and damaging 370,000 homes. More precisely, the study analyzed tweets related to seven confirmed truths and seven false rumors with from 42 to 700 unique messages per case. It reports the results of a *content analysis* (Stemler 2001), which identified the truth-affirming, truth-denying, and question-posing tweets. As many as 96% of the tweets related to confirmed truths got judged as truth-affirming, less than 4% as

³⁸ Health & Wellness Yahoo! Groups (<http://groups.yahoo.com/neo/dir/1600060813>)

³⁹ The aforementioned March 2012 rate (Twitter Blog 2012)

questions, and as little as 4% as truth-denying. In contrast, of the tweets related to false rumors, as many as 38% were seen as truth-denying, more than 17% as questions, and only 45% as truth-affirming. That is, false rumors were questioned much more than confirmed truths. This gives evidence for the large proportion of questions in tweet message chains indicating unreliable information.

6.10 Future work

It is clear that social media has become an important channel of communication for everyone, the public, organizations and governments. Social media – the information it contains and the interactions that it enables – can be mined to benefit health care in a variety of ways, for both laypersons and professionals. We have described some systems that exploit social media to identify and monitor health related crisis, to discover adverse side effects of drugs, or to monitor specific topics.

This is only the beginning, as we are most likely to see new applications in the future. The CSIRO work in this area, in addition to what we describe above, has several other research foci. For example, we are looking at attitudes towards depression and suicide in social media, by trying to identify how people react to social media posts that express distress. We are also exploring how people talk about diet-related issues, to see if we can obtain from social media some valuable insights that are similar to those that can be obtained through surveys and interviews. We are also looking at how people might interact and influence each other in healthy living-related online communities. We are seeking to understand how people discuss various topics on social media (Paris et al. 2012) and to learn more about the impact of such communications on medicine, health care, and patients themselves.

In all this work, we have to address the issues mentioned earlier in the chapter, in particular the quantity (and speed) of the information, the noise it contains, and the potential relationships that exist (both within one channel, and across many media channels). In addition, the language employed on social media can be challenging for NLP techniques. Indeed, the language found on social media is usually informal, and often it contains unconventional (and at times colorful) vocabulary, syntax, and punctuations – sometimes through misspellings and sometimes by choice, expressing the creativity of the writer.

Our overall aim is to develop a set of tools to help laypersons, professionals, and public health organizations find accurate and reliable information from social media for a variety of purposes and applications as well as to evaluate them in real tasks. The tools include text classifiers, topic detectors, document summarizers, visualization mechanisms, and social network analyzers.

Acknowledgments

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. NICTA is also funded and supported by the Australian Capital Territory, the New South Wales, Queensland and Victorian Governments, the Australian National University (ANU), the University of New South Wales, the University of Melbourne, the University of Queensland, the University of Sydney, Griffith University, Queensland University of Technology, Monash University and other university partners.

We are thankful for our co-supervised students' contribution at NICTA. We express our gratitude to our 3rd year engineering student Xing Yu (Frank) Su from ANU for his contribution to developing and evaluating the TweetDetector search engine during his summer scholarship at NICTA in 2010–2011. We thank Frank for continuing this work as his honors project during his 4th year in 2011 together with his team of 3rd year engineering students Kalinga Hulugalle, Peter Walton, and Riley Kidd and mentor David Needham in ANU. We acknowledge the contribution by PhD students Aapo Immonen from the University of Eastern Finland and Karl Kreiner from the AIT Austrian Institute of Technology GmbH to developing and evaluating the NICTA automated classifier for crisis managers during their visits at NICTA in 2011 and 2013.

We acknowledge the contribution of the NICTA team of researchers and engineers behind EventWatch. In particular, we thank Scott Sanner and Wray Buntine from NICTA and ANU for leading this scientific research and development.

For the CSIRO work, we acknowledge the contributions of the CSIRO staff, researchers, and engineers, responsible for the ESA and Vizie prototypes. For ESA, we acknowledge (in alphabetical order) Mark Cameron, Jessie Yin, Andrew Lampert, Sarvnaz Karimi, Robert Power, and Bella Robinson. For the Vizie prototype, we would like to acknowledge in particular James McHugh, Brian Jin, Payam Aghaei Pour and Hassan Asghar for their software engineering work on the prototype, under the guidance of the lead researcher, Stephen Wan. Finally, for both systems, we would like to thank John Colton and all our users for helping us shape the tools and their organizations for contributing to the funding of the research and development towards the tools.

References

- Ahlqvist, T., Bäck, A., Halonen, M. & Heinonen, S. (2008) 'Social media road maps. Exploring the futures triggered by social media', *VTT Tiedotteita – Res Notes*, 2454:78.
- Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D., Harper, D. J., Hiemstra, D., Hofmann, T., Hovy, E., Kraaij, W., Lafferty, J., Lavrenko, V.,

- Lewis, D., Liddy, L., Manmatha, R., McCallum, A., Ponte, J., Prager, J., Radev, D., Resnik, P., Robertson, S., Rosenfeld, R., Roukos, S., Sanderson, M., Schwartz, R., Singhal, A., Smeaton, A., Turtle, H., Voorhees, E., Weischedel, R., Xu, J. & Zhai, C. (2003) 'Challenges in information retrieval and language modeling: report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002', *SIGIR Forum*, 37(1):31–47.
- Bates, D. W., Evans, R. S., Murff, H., Stetson, P. D., Pizziferri, L. & Hripsak, G. (2003) 'Detecting adverse events using information technology', *JAMA*, 10(2):115–128.
- Bates, D. W., Cullen, D. J., Laird, N., Petersen, L. A., Small, S. D., Servi, D., Laffel, G., Sweitzer, B. J., Shea, B. F., Hallisey, R., Vander Vliet, M., Nemeskal, R., Leape, L. L., Bates, D., Hojnowski-Diaz, P., Petrycki, S., Cotungo, M., Patterson, H., Hickey, M., Kleefeld, S., Cooper, J., Kinneally, E., Demonaco, H. J., Dempsey Clapp, M., Gallivan, T., Ives, J., Porter, K., Thompson, B. T., Hackman, J. R. & Edmondson, A. (1995) 'ADE prevention study group. Incidence of adverse drug events and potential adverse drug events: implications for prevention', *JAMA*, 274(1):29–34.
- Bates, D. W., Spell, N., Cullen, D. J., Burdick, E., Laird, N., Petersen, L. A., Small, S. D., Switzer, B. J. & Leape, L. L. (1997) 'The costs of adverse drug events in hospitalized patients. Adverse drug events prevention study group', *JAMA*, 277(4):307–311.
- Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., Leonard, C. E. & Holmes, J. H. (2011) 'Identifying potential adverse effects using the Web: a new approach to medical hypothesis generation', *J Biomed Inform*, 44(6):989–1006.
- Bian, H., Topaloglu, U. & Yu, F. (2012) 'Towards large-scale Twitter mining for drug-related adverse events'. In Yang, C. C., Chen, H., Wactlar, H., Combi, C. & Tang, X. (program chairs). *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing (SHB '12)*. New York, NY, USA: ACM.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) 'Latent Dirichlet allocation', *J Mach Learn Res*, 3:993–1022.
- Bruns, A., Burgess, J., Crawford, K. & Shaw, F. (2012) *#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods*. Brisbane, QLD, Australia: ARC Centre of Excellence for Creative Industries and Innovation.
- Cameron, M., Power, R., Robinson, B. & Yin, J. (2012) Emergency Situation Awareness from Twitter for Crisis Management. In Mille, A., Gandon, F., Misselis, J., Rabinovich, M. & Staab, S. (general and program chairs). *Proceedings of the 21st International Conference Companion on World Wide Web (WWW '12 Companion)*. New York, NY: ACM.
- CampaignBrief. (2013) *Magna Global Research Reveals Australian Mobile Internet Access Has Risen by 208% in Three Years*. [Online] 8 April Available from: <http://www.campaignbrief.com/2013/04/magna-global-research-reveals.html>. [Accessed: 15 Dec 2013].
- Can, F. (1993) 'Incremental clustering for dynamic information processing', *ACM Trans Inform Process Systems*, 1(2):143–164.
- Cataldi, M., Di Caro, L. & Schifanella, C. (2010) 'Emerging topic detection on Twitter based on temporal and social terms evaluation'. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining (MDMKDD '10)*. New York, NY: ACM.
- Chee, B. W., Berlin, R. & Schatz, B. (2011) 'Predicting adverse drug events from personal health messages', *AMIA Annu Symp Proc*, 2011:217–226.
- Cohen, J. (2011) *History's Worst Nuclear Disasters, History in the Headlines*. [Online]. 18 March Available from: <http://www.history.com/news/historys-worst-nuclear-disasters>. Accessed: 15 Dec 2013].

- Coiera, E. (2003) *Guide to Health Informatics*. 2nd edition. London, UK: Arnold Publication.
- Colineau, N. & Paris, C. (2010) 'Talking about your health to strangers: understanding the use of online social networks by patients', *New Rev Hypermed Multimed*, 16(1–2):141–160.
- Dua, S., Acharya, U. R. & Dua, P., (eds.) (2014) *Machine Learning in Healthcare Informatics. Springer Intelligent Systems Reference Library*. 56. Heidelberg, Germany: Springer.
- Dusc, The Drug Utilisation Sub-Committee. (2000) *Australian Statistics on Medicines*. Canberra, ACT, Australia: Department of Health and Aging, Australia.
- Experian Hitwise. (2008) *Google Receives 87.81 Percent of Australian Searches in June 2008*. [Online]. Available from: <http://www.hitwise.com/au/press-centre/press-releases/2008/ap-google-searches-for-june/>. [Accessed: 15 Dec 2013].
- Fox, S. (2011) *Health Topics: 80% of Internet Users Look for Health Information Online. Technical report, Pew Research Center*. [Online] 1 February. Available from: <http://www.pewinternet.org/Reports/2011/HealthTopics.aspx>. [Accessed: 15 Dec 2013].
- Fox, S. & Jones, S. (2009) *The Social Life of Health Information. Pew Internet & American Life Project*. [Online] 11 June. Available from: <http://www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx>. [Accessed: 15 Dec 2013].
- Giacomini, K. M., Krauss, R. M., Roden, D. M., Eichelbaum, M., Hayden, M.R. & Nakamura, Y. (2007) 'When good drugs go bad', *Nature*, 446(7139):975–977.
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. & Smith, N. A. (2011) Part-of-speech tagging for Twitter: annotation, features, and experiments. In Lin, D. (general chair). *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers – Volume 2 (HLT '11)*. Stroudsburg, PA: ACL.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2008) 'Detecting influenza epidemics using search engine query data', *Nature*, 457:1012–1014.
- Griffen, G., Jones, R. & Paris, C. (2012) Strategic implications of social media for emergency management. In Clark, M. & Griffen, G. (eds.). *Next Generation Disaster and Security Management*. Canberra, ACT, Australia: The Australian Security Research Council.
- Gupta, A., Lamba, H. & Kumaraguru, P. (2013a) \$1.00 per RT #BostonMarathon #PrayForBoston: analyzing fake content on Twitter. In *Proceedings of the eCrime Researchers Summit (eCRS) 2013*. San Francisco, CA, USA, September 2013.
- Gupta, A., Lamba, H. & Kumaraguru, P. (2013b) Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In Schwabe, D., Almeida, V., Glaser, H., Baeza-Yates, R. & Kaist, S. M. (general and program chairs). *Proceedings of the 22nd International Conference on World Wide Web Companion (WWW '13 Companion)*. New York, NY: ACM.
- Hagar, C. (2006) Using research to aid the design of a crisis information management course. In *Proceedings of ALISE SIG Multicultural, Ethnic & Humanistic Concerns (MEH)*.
- Hamm, M. P., Chrisholm, A., Shulhan, J., Milne, A., Scott, S. D., Klassen, T. P. & Hartling, L. (2013) 'Social media use by health care professionals and trainees: a scoping review', *Acad Med*, 88(9):1376–1383.
- Heinzelman, J. & Waters, C. (2010) *Crowdsourcing Crisis Information in Disaster-affected Haiti*. United States Institute of Peace Special Report. Washington DC: United States Institute of Peace.
- Karimi, S., Yin, J. & Paris, C (2013) Classifying microblogs for disasters. In Culpepper, S., Zuccon, G. & Sitbon, L. (eds.). *Proceedings of the 18th Australasian Document Computing Symposium (ADCS '13)*. New York, NY: ACM.

- Kireyev, K., Palen, L. & Anderson, K. M. (2009) Applications of topics models to analysis of disaster-related Twitter data. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*. Amherst, MA, December 2009.
- Kobayashi, M. & Takeda, K. (2000) 'Information retrieval on the Web', *ACM Comput Surv*, 32(2):144–173.
- Kreiner, K., Immonen, A. & Suominen, H. (2013) Crisis management knowledge from social media. In Culpepper, S., Zuccon, G. & Sitbon, L. (eds.). *Proceedings of the 18th Australasian Document Computing Symposium (ADCS '13)*. New York, NY: ACM.
- Kummervold, P., Chronaki, C., Lausen, B., Prokosch, H., Rasmussen, J., Santana, S., Staniszewski, A. & Wangberg, S. (2008) 'eHealth trends in Europe 2005–2007: a population-based survey', *J Med Internet Res*, 10(4):e42.
- Kupiec, J., Pedersen, J. & Chen, F. (1995) A trainable document summarizer. In Fox, E. A., Imgwersen, P. & Fidel, R. (eds.). *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95)*. New York, NY: ACM.
- Laub, K. (2009) *Libyan estimate: at least 30,000 died in the war*. *The Guardian* [Online] 8 September. Available from: <http://www.theguardian.com/world/feedarticle/9835879>. [Accessed: 15 Dec 2013]
- Lazarou, J., Pomeranz, B. H. & Corey, P. N. (1998) 'Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies', *JAMA*, 279(15):1200–1205.
- Leaman, R., Wojtulewicz, L., Sullivan, R., Sakariah, A., Yang, J. & Gonzalez, G. (2010) 'Towards Internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks'. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP '10)*. Stroudsburg, PA: ACL.
- Lim, K. W., Chen, C. & Buntine, W. 2013 'Twitter-network topic model: a full Bayesian treatment for social network and text modeling'. In *Proceedings of NIPS 2013 Topic Model Workshop*. Lake Tahoe, NV, USA, December.
- Liu, J., Li, A. & Seneff, S. (2011) 'Automatic drug side effect discovery from online patient-submitted reviews: focus on statin drugs'. In *Proceedings of the First International Conference on Advances in Information Mining and Management (IMMM '11)*. Barcelona, Spain, October 2011.
- Madoka (2013) *Japan Breaks Record for Tweets per Second again with 'BALS'*. *japanCRUSH*. [online] 5 August. Available from: <http://www.japancrush.com/2013/stories/japan-breaks-record-for-tweets-per-second-again-with-bals.html>. [Accessed: 15 Dec 2013].
- Manning, C. D. & Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. London, England: The MIT Press.
- Manning, C. D., Raghavan, P. & Schütze, H. (2008) *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
- Mehrotra, R., Sanner, S., Buntine, W. & Xie, L. (2013) Improving LDA topic models for microblogs via Tweet pooling and automatic labeling. In Jones, G. J. F., Seridan, P., Kelly, D., de Rijke, M. & Sakai, T. (general and program chairs). *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. New York, NY: ACM.
- Mendoza, M., Poblete, B. & Castillo, C. (2010) Twitter under crisis: can we trust what we RT? In Melville, P., Leskovec, J. & Provost, F. (conference chairs). *Proceedings of the 1st Workshop on Social Media Analytics (SOMA '10)*. Washington, DC: ACM.
- Moore, T. J., Cohen, M. R. & Furberg, C. D. (2007) 'Serious adverse drug events reported to the Food and Drug Administration, 1998–2005', *Arch Intern Med*, 167(16):1752–1759.

- Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J.K., Irwin, A. & Hoving, C. (2013) 'A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication', *J Med Internet Res*, 15(4):e85.
- Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011) 'Natural language processing: an introduction', *J Am Med Inform Assoc*, 18(5):544–551.
- Paris, C., Thomas, P. & Wan, S. (2012) 'Differences in language and style between two social media communities'. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM '12)*. Dublin, Ireland: AAAI Press.
- Paris, C. & Wan, S. (2011) Listening to the community: social media monitoring tasks for improving government services. In Tan, D., Begole, B. & Kellogg, W. A. (general and program chairs). *Proceedings of the HI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. New York, NY: ACM.
- Power, R., Robinson, B. & Radcliffe, D. (2013) Finding fires with Twitter. In Karimi S & Verspoor K. *Proceedings of the 2013 Australasian Language Technology Association Workshop (ALTA '13)*, Brisbane, QLD, Australia, December 2013.
- Price Waterhouse Coopers (2010) *Health Leaders Media Breakthroughs: The Impact of Personalized Medicine Today*. [Online] April. Available from: <http://www.healthleadersmedia.com/breakthroughs/250079/The-Impact-of-Personalized-Medicine-Today>. [Accessed: 15 Dec 2013]
- Radev, D., Otterbacher, J. & H Qi, D. T. (2003) 'MEAD ReDUCs: Michigan at DUC 2003'. In. *Proceedings of the Document Understanding Conference 2003: Workshop on Text Summarization*. Edmonton, Canada, May 2003.
- Sakaki, T., Okazaki, M. & Matsuo, Y. (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In Rappa, M., Jones, P., Freire, J. & Chakrabarti, J. (general and program chairs). *Proceedings of the 19th International Conference Companion on World Wide Web (WWW '10 Companion)*. New York, NY: ACM.
- Signorini, A., Segre, A. M. & Polgreen, P. M. (2011) 'Public concern in the U. S. during the influenza AH1N1 pandemic', *PLoS ONE*, 6(5):e19467.
- Stemler, S. (2001) 'An overview of content analysis', *Pract Assess Res Eval*, 7(17):1–9.
- Su, X. Y., Kidd, R., Hulugalle, K., Walton, P., Suominen, H., Hanlen, L. & Needham, D. (2011a) *Tweet Detector Project, Final Presentation*. The Australian National University, College of Engineering and Computer Science, Canberra, ACT, Australia, 19 October.
- Su, F., Suominen, H. & Hanlen, L. (2011b) 'Machine intelligence for health information: capturing concepts & trends in social media via query expansion', *Stud Health Technol Inform*, 168:150–157.
- Sullivan, D. (2010) *Where have all the old Tweets gone? Search Engine Land*. [Online] 14 January. Available from: <http://searchengineland.com/where-have-all-the-old-tweets-gone-33579>. [Accessed: 15 Dec 2013].
- Suominen, H., Pahikkala, T. & Salakoski, T. (2008) Critical points in assessing learning performance via cross-validation. In Honkela, T., Pöllä, M., Paukkeri, M.-S. & Simula, O. (eds.). *Proceedings of the 2nd International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR '08)*. Porvoo, Finland, 17–19 September 2008.
- Suominen, H. & Salakoski, T. (2010) 'Supporting communication and decision making in Finnish intensive care with language technology', *J Healthcare Eng*, 1:595–614.

- Sutton, J., Palen, L. & Shklovski, I. (2008) Backchannels on the front lines: emergent use of social media in the 2007 Southern California wildfires. In Fiedrich, F. & Van de Walle, B. (eds.). *Proceedings of Proceedings of the 5th International Information Systems for Crisis Response and Management Conference (ISCRAM '08)*. Workshop, Washington DC, USA, May 2008.
- The Australian (2013) *Twitter 'Can't Be Trusted in a Crisis'*. [Online] 2 December. Available from: <http://www.theaustralian.com.au/media/twitter-cant-be-trusted-in-a-crisis/story-e6frg996-1226772527245#>. [Accessed: 15 Dec 2013].
- The Telegraph (2009) *New York Plane Crash: Twitter Breaks the News, again*. [Online] 16 January. Available from: <http://www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html>. [Accessed: 15 Dec 2013].
- Twitter Blog (2011) *#numbers, 200 Million Tweets per Day, and The Engineering behind Twitter's New Search Experience*. [Online]. 14 March, 30 June, 31 March. Available from: <https://blog.twitter.com/2011/numbers>, <https://blog.twitter.com/2011/200-million-tweets-day>, and <https://blog.twitter.com/2011/engineering-behind-twitter%E2%80%99s-new-search-experience>. [Accessed: 15 Dec 2013].
- Twitter Blog (2012) *Twitter Turns Six*. [Online]. 21 March Available from: <https://blog.twitter.com/2012/twitter-turns-six>. [Accessed: 15 Dec 2013].
- Twitter Blog (2013) *Celebrating #Twitter7 and New Tweets per Second Record, and How!* [Online]. 21 March, 16 August. Available from: <https://blog.twitter.com/2013/celebrating-twitter7> and <https://blog.twitter.com/2013/new-tweets-per-second-record-and-how>. [Accessed: 15 Dec 2013].
- The World Bank (2013) *Internet Users (per 100 People)*. [Online]. Available from: <http://data.worldbank.org/indicator/IT.NET.USER.P2>. [Accessed: 15 Dec 2013].
- Wan, S. & McKeown, K. (2004) 'Generating overview summaries of ongoing email thread discussions'. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*. Stroudsburg, PA: ACL.
- Wan S. & Paris, C. (2014) 'Improving government services with social media feedback'. In *Proceedings of the 2014 International Conference on Intelligent User Interfaces (IUI '2014)*. Haifa, Isreal, Feb 24–27, 2014.
- Wu H. & Stanhope, S. J. (2012) An early warning system for unrecognized drug side effects discovery. In Mille, A., Gandon, F., Misselis, J., Rabinovich, M. & Staab, S. (general and program chairs). *Proceedings of the 21st International Conference Companion on World Wide Web (WWW '12 Companion)*. New York, NY: ACM.
- Yin, J., Lampert, A., Cameron, M., Robinson, B. & Power, R. (2012) 'Using social media to enhance emergency situation awareness', *IEEE Intell Syst*, 27(6):52–59.
- Young, A. & Bloor, J. (2009) *Medical Twitter. BMJ Careers* [Online] 10 June. Available from: <http://careers.bmj.com/careers/advice/view-article.html?id#equal#20000214>. [Accessed: 15 Dec 2013].

**Part III Using speech and audio technologies
for improving access to online content for
the computer-illiterate and the visually
impaired**

Olufemi Oyelami

7 An empirical study of user satisfaction with a health dialogue system designed for the Nigerian low-literate, computer-illiterate, and visually impaired

Abstract: The advent of the Internet has made this elaborate communication network a repository for many different kinds of information sharing. Among the information normally searched for on the Internet is health-based information. This information helps the consumers to carry out “self-care” (activities that contribute to maintaining a state of well-being); in so doing, it empowers them to initiate lifestyle changes and new regimens to maintain good health. However, this information on the Internet is primarily delivered in text format. Due to inadequate Internet access and the low level of literacy in Nigeria, vital health information is only available to a small percentage of the population. This chapter reports on the development, acceptability, and user satisfaction with a dialogue system providing health information about lassa fever, malaria fever, typhoid fever and yellow fever. This system caters to the needs of those who lack Internet access or who are computer-illiterate, low-literate, or visually impaired. A cross-sectional study was conducted using a questionnaire that gathered demographic data about the study participants and their satisfaction and readiness to accept the system. The user satisfaction results showed a mean of 3.98 (approximately 4), which is the recommended average for a good usability study. Dialogue systems of this kind help to provide cost-effective and equitable access to health information that can protect the population from tropical disease outbreaks. They serve the low-literate, the computer-illiterate, and the visually impaired.

7.1 Introduction

The Internet serves as a repository for different kinds of information. There is rarely a subject on which information cannot be gleaned on the Internet, including health. There were 2.1 billion Internet users world-wide in 2011 (Hadlaczky et al. 2013). Jesaimini et al. (2013) state: “One of the most cited reasons for accessing the Internet is searching for health information.” Patients access medical information

through taking part in online discussion groups, searching for health information in medical databases, arranging for consultations with doctors, and using self-administered treatment and diagnosing tools found on line (Bessel et al. 2002). Patients also seek, in addition to health information, emotional support through websites that are devoted to a particular disease as well as from online support groups and electronic mailing lists that give newsy updates to patients who have signed up for these regular email alerts (Alejandro & Gagliardi 1998).

Access to useful health information (HI) and medical information (MI) on the Internet can be lifesaving in underdeveloped countries. For example, in Nigeria the current life expectancy is 49 years according to the 2010 report of the World Health Organization (WHO). That report lists malaria, diarrhea, pneumonia, prematurity, birth asphyxia, neonatal sepsis, HIV/AIDS, congenital abnormalities and injuries as the most common causes of death, in rank order. Some of the major illnesses that lead to mortality in Nigeria, such as those listed above, could have been prevented with simple medications and healthy lifestyles (Acho 2005). Most certainly, the situation could have been different if the populace were aware of the availability of health information on the Internet and were able to make proper use of such information. This is especially so because the Internet has a lot of prospects in supporting health care and self-care (Bernhardt 2000; Baker et al. 2003; Lintonen et al. 2008; Weaver et al. 2009). Health-related information on the Internet could help patients to be better informed, obtain more knowledge about their illnesses and conditions, and consequently be more involved in the decision making process concerning their health, rather than passively sitting by as the peril of illness and disease consume their vitality. No doubt, such access to online health information could improve patients' health by ensuring that they have more appropriate healthcare services that are sorely needed at the early stages of illness (Diaz et al. 2002; Lupiáñez-Villanueva 2011).

7.2 Related work

Jesaimini et al. (2013) reviewed literature up to March 2012 in an effort to investigate which kinds of users searched for health information on the Internet and for what purposes. Their specific focus was on medication information in particular. Their study covered patients in the general population rather than in one specific region. They studied patients' use of the Internet in North America and Europe, in the Middle East and Asia, and in Australia and New Zealand. Their results showed that nearly one half of the general population and 50%–99% of adults suffering from a chronic disease had used the Internet to search for health information, primarily about a specific disease, its treatment, exercise, and diet.

As regards medications, approximately one half of the online health information seekers, whether patients or not, looked for medical information concerning side effects, drug safety, interactions, update on drugs currently consumed, new drugs, and over-the-counter or alternate medications. Women, adults older than 50 years, and well-educated people searched considerably more frequently for health information and medical information. The reasons to search on line for medical information were convenience, broad range of information, and peers' opinions. The online searches for medical information did not replace health professionals' information, but offered additional information and a possibility to crosscheck. Interestingly enough, the study results also showed that not only can online medical information reassure a patient or incentivize them to ask questions from the treating physician, but online health-related information can likewise confuse the patient. This is so because the patients' lack of medical expertise creates confusion when presented with large amounts of, sometimes contradictory, information which can be hard to sift through and digest.

Oyelami et al. (2013) investigated Nigerians' Internet pattern usage. This included studying patients' awareness of the availability of health-related information on line, as well as an examination of the key factors that influenced their use of the Internet for self-care health information. A questionnaire-based assessment of 205 individuals selected randomly was carried out. The results indicated that 61% of the participants used the Internet for *self-care* (self-diagnosis and treatment in lieu of seeking help from the medical profession) and were aware of the availability of health information on the Internet, which they readily pursued to find answers to their health concerns. The participants in this study also reported that they had used the Internet for purposes other than seeking health-related information. Those purposes included communication, social networking, general research, and banking. The results validated study participants' perceived ease of use, *compatibility* (consistency with the values or norms of potential adopters of a technology and the similarity with existing standards), Internet *self-efficacy* (belief in a person's ability to succeed in a specific situation), and technical support and training, as factors to consider in using the Internet for self-care.

Jo et al. (2010) carried out a survey to reveal the patterns of utilization of health information on the Internet among native residents of the metropolitan city of Incheon and simultaneously in the Gangwon province of South Korea. Their results revealed the following categorical breakdown for the health information that people sought on the Internet: general health tips (64.2%); disease specific information (32.0%); shopping for health commodities such as HIV test kits, contraceptive devices, etc. (23.7%); and selection of hospitals (19.3%). The survey showed that people with a higher education and higher income level were

inclined to use the Internet more often for health information than those who were less educated and had less income. Similarly, metropolitan city residents used health information found on the Internet more often than those leading a more humble lifestyle living in the outer lying province. One's personal health status appeared to be the most important factor in determining the use of the Internet for searching information about general health tips. For example, healthy people (68.3%) used the Internet more than those plagued by illnesses (44.4%). However, among the population of ill people who availed themselves of the Internet, they were found to use the Internet most frequently for disease-specific information (62.6%). Residence area (where a person resides) was the most important factor of online shopping for health commodities. For instance, whereas 31.8% of city dwellers used the Internet for purchasing health commodities, only 19% of those living in the province used it for the same purpose. Similarly, residence area, age, and health examination were the determinant factors for the utilization of the Internet for hospital selection.

AlGhamdi & Moussa (2012) through a *self-administered questionnaire* (a questionnaire that is administered without an interviewer) carried out a study to determine how the public uses the Internet in Saudi Arabia to search for health-related information. As part of the study, the respondents were asked to evaluate their perceptions of the quality of information they found on the Internet when compared with the information they obtained from their own healthcare providers. Their results showed that 87.8% of the study respondents used the Internet generally and 58.4% used the Internet for searching specifically for health-related information. While 89.3% reported that a doctor was their primary source of health information, 84.2% of those surveyed agreed that searching for health information on the Internet was useful. The reasons given were: (1) curiosity (92.7%); (2) not getting sufficient information from their doctors (58.5%); and (3) not trusting the information given to them by their doctors (28.2%). In fact, 44% of study participants searched for health information before going to the clinic; 72.5% of the study respondents discussed the information they obtained from the Internet with their doctors. And nearly all those who did so (71.7%) believed that this positively affected their relationship with their doctor. Health information search was more frequent among the 30–39 year age group as well as those with university or higher education, employed individuals, and high-income groups.

The work by Sadasivam et al. (2013) reveal that not only do individuals who need health information search for it, but there is similarly a category of health information seekers called “surrogate seekers” – those who actively search the Internet for relevant health information for persons other than themselves. Members of this category search for health-related information for their family members or friends who may be suffering from serious health conditions. It is

important to address this category because they are frequent visitors of the Internet along with those who search the web for information related to their own conditions. The study seeks to identify the unique characteristics of surrogate seekers, showing how they differ from self-seekers of health information. The researchers contend that by “identifying the unique characteristics of surrogate seekers [this] would help in developing Internet interventions that better support these information seekers.”

From all the related work described above, it can be seen that the vital health information accessed by the patient, or their advocate, appears in text format only. However, if a consumer is not computer literate, has no access to the Internet, is visually impaired, or is suffering from literacy problems in general, such a consumer is clearly at a disadvantage. Hence, it is imperative that we attend to the needs of these special categories of individuals who would benefit from the massive amount of health-related information found on line.

7.3 Dialogue systems

According to Bickmore & Giorgino (2006), a dialogue is discourse between two or more parties, including a human and a computer. Bickmore and Giorgino (2006) and Alan et al. (2004) defined a dialogue system as a computer system that communicates with a human.

Even though there are proprietary solutions for developing dialogue systems, VoiceXML is the W3C standard designed for human-computer audio dialogues that feature synthesized speech, digitized audio, recognition of spoken, DTMF (Dual Tone Multi Frequency) key input, recording of spoken input, telephony and mixed initiative conversations (José 2007; W3C 2001). The main goal of VoiceXML is to bring the full power of Web development and content delivery to voice response applications and to free the authors of such applications from low-level programming and resource management. VoiceXML enables integration of voice services with both data services, using the familiar client-server paradigm (W3C 2001). For a traditional webpage, a Web browser will make a request to a Web server, which will, in turn, send an HTML document to the browser to be displayed visually to the user. However, for a dialogue system, it is the VoiceXML Interpreter that sends the request to the Web server, which will return a VoiceXML document to be presented as a dialogue system via a telephone.

Dialogue systems for calling up web-based medical content play a special role in underdeveloped countries. Nigeria presents an interesting test case in that the low level of computer literacy serves as one of the major impediments to accessing health information on the Internet (Jegade & Owolabi 2003;

Esharenana & Emperor 2010). Furthermore, only 27.3% of the world population has access to computers, and 25.9% have access to the Internet (ITU 2009). Consequently, one technology that can be used to overcome the accessibility problem is telephonic communication. In fact, telephones outnumber computers on the planet (José 2007), and 67% of the world population has access to them as stated in the ITU report (2009). Given the ubiquity of telephones, HI and MI may be made accessible to the Nigerian populations of the computer-illiterate, low-literate and the visually impaired via the proper use of the spoken dialogue system, which we describe below. This system is accessible via both fixed lines (landlines) and mobile phones. Dialogue systems provide interactive voice dialogues between a human and a computer. They have the potential of being used to provide ubiquitous, cost-effective and wide-scale services to a vast majority of people (David 2006). They can also be used to provide access to health information available on the Internet to the visually-impaired. In this work, below, we explore how Health Dialogue System (HDS) provides access to health information. We show how the system was developed and evaluate its performance for acceptability and user-satisfaction.

7.4 Methods

The Health Dialogue System (HDS) was developed using VoiceObjects Desktop for Eclipse 11. This is an Eclipse-based IDE for designing, developing, testing, deploying and administering of voice, video, text and Web-based applications. Voxeo Prophecy 13 was used as the implementation platform consisting of a speech server and VoiceXML engine. Voxeo Prophecy is a standards-based premise voice platform that is used by Voxeo customers worldwide for inbound IVR, outbound notification, innovative VoIP applications, and more. All these tools allow for testing of voice applications without having to deploy them on the telecommunication service providers' networks. A white female voice was used by HDS in interacting with the participants. By default, Voxeo Prophecy's text-to-speech (TTS) system can be used in either a white female voice or a white male voice mode. HDS was tested using an in-built soft phone in Voxeo Prophecy.

7.4.1 Participants

The evaluation of HDS was carried out among 19 undergraduates of Landmark University, which is located in Omu-Aran, Kwara State, Nigeria. Of all of the 19 subjects that participated in the study, one subject did not complete the questionnaire. Each of the subjects that participated in the evaluation was informed beforehand of what services HDS offers and how to interact with the system. Each was subsequently

invited to test the system on a laptop computer running Windows 8. After the test, each subject was given a questionnaire to fill out.

7.4.2 Demographics of the participants

Nine of the participants were male while five were females. The remaining four subjects chose not to specify their sex. Ten of the participants were less or equal to 20 years of age, five were within the 21–30 age range, while three did not specify their age range.

7.4.3 Data collection

In measuring user satisfaction of the system, items in questionnaires used in similar studies by Kwan & Jennifer (2005) and Walker et al. (1999) were adopted. The measures used in the questionnaires have both face and content validities. For face validity, all measures were constructed by experts with over 10 years of experience in usability tests of mobile and speech user interface (SUI) applications. In terms of content validity, the measures covered all dimensions of usability in telephony applications as defined by the European Telecommunications Standard Institute (ETSI) (Kwan & Jennifer 2005). However, a modification was made to the questionnaire by Kwan & Jennifer (2005) by the changing of some adjectives to their simpler synonyms in a bid to aid the participants' understanding. The questionnaire used was scaled 1–5.

7.4.4 Data analysis

The completed questionnaires by the participants were analyzed using Microsoft Excel 2010. Descriptive statistics-frequencies and percentages were calculated.

7.5 Health dialogue system (HDS)

The prototype system developed provides health information about fevers rampant in Nigeria. The system when called up by the user proceeds to welcome the user and then quickly informs the user of the services it renders. The user is then expected to make a choice from a list of diseases – lassa fever, malaria fever, yellow fever and typhoid fever – in order to get information about those diseases. Once a selection is made, the system presents general information about the specific fever. The caller is then asked to select information about any of the following aspects of the fever:

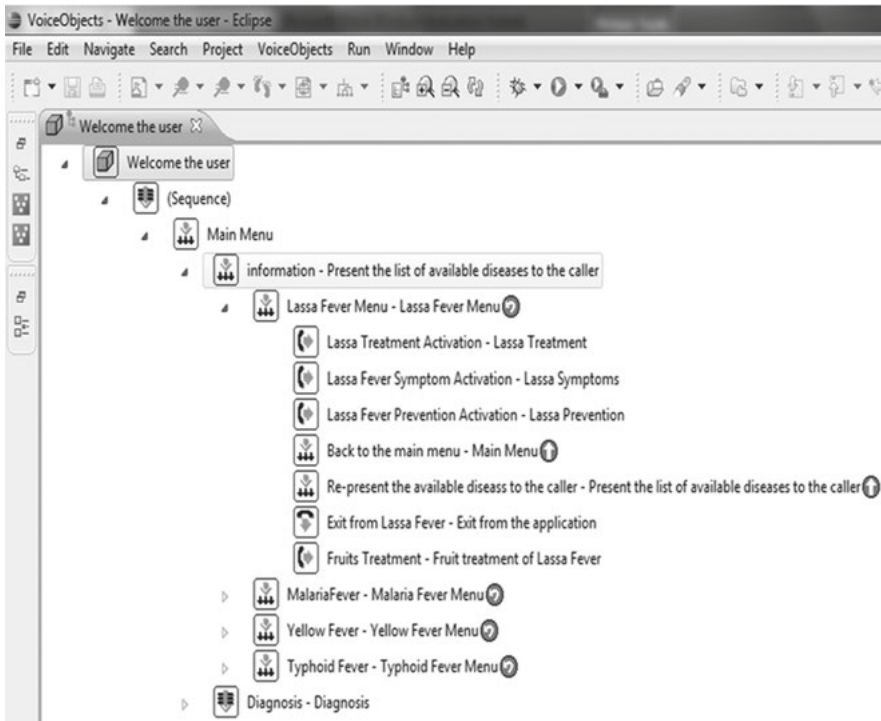


Fig. 7.1: Dialogue flow of HDS.

symptoms of the fever; how to treat it; information about fruits that can be used in treating the fever; and how to prevent it. The caller also has the option of listening to the list of the diseases again before exiting from the application. The dialogue flow generated by VoiceObjects is shown above in Fig. 7.1.

Figure 7.2 below shows HDS being tested with the Prophecy 13 in-built soft phone.

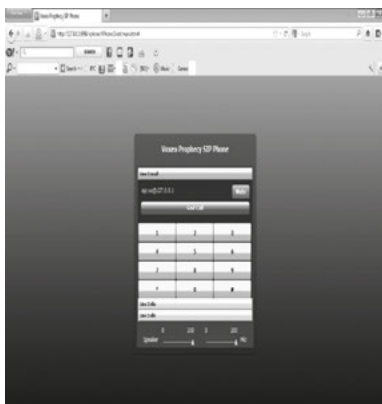


Fig. 7.2: Calling HDS with Prophecy 13 in-built soft phone.

7.6 Results

7.6.1 Experiences with mobile/computing devices

As shown in Tab. 7.1 below, five (27.8%) of the study participants rated their experiences/skill in the use of computer as “expert.” Ten (55.6%) rated their skills as “good.” Two (11%) rated their skills as “average” while one (5.6%) rated his computer skills as “novice.” Eight of the participants (44.4%) primarily used a desktop PC to do their work, while 10 (55.6%) made their laptop their primary computing device. Fifteen of the participants (83.3%) owned a mobile phone or personal digital assistant (PDA); two (11.1%) did not own either a mobile phone or a PDA. Lastly, one participant (5.6%) did not specify this option about a mobile device all together. Of all those that owned a mobile device, 14 (77.8%) had owned a mobile phone/PDA for more than 2 years while only one (5.6%) participant had owned such a device for 2 years or less. Three of the participants (16.6%) did not specify the duration of their ownership of a mobile device. When questioned about frequency of usage, as seen in Tab. 7.1, the results show that 14 (77.8%) of the respondents made or received calls more than seven times a week; one respondent (5.6%) made calls 5–6 times in a week; one respondent

Tab. 7.1: Experiences with mobile/computing devices.

Item	Category	%
Software usage skill	Novice	5.6
	Average	11
	Good	55.6
	Expert	27.8
Device used for work	Laptop	55.6
	Desktop PC	44.4
Ownership of phone/PDA	Mobile phone/PDA	83.3
	No ownership	11.1
	No response	5.6
Duration of ownership of phone/PDA	> 2 years	77.8
	< 2 years	5.6
	No response	16.6
Frequency of making or receiving of calls weekly	> 7 times	77.8
	5–6 times	5.6
	1 time	5.6
	No response	11

(5.6%) made calls once a week; and two of the respondents (11%) did not make any calls at all.

7.6.2 User satisfaction and acceptability of HDS

In response to the question, “**Would you like to access health information using this kind of system? If yes, why?**” Seventeen of the respondents (94%) responded “yes” while one (6%) did not specify this option. These results showed a user satisfaction mean average of 3.98. Such results imply that all the participants approved access to health information via dialogue systems in that a satisfaction result of 3.98, which is approximately 4, comports with the recommended average for good usability studies on 1–5 scale (Sauro & Kindlund 2005).

Table 7.2 below lists all the reasons given by the study participants for their satisfaction and acceptance of HDS.

Tab. 7.2: Satisfaction of use and acceptability of HDS.

“It can be used in times of emergencies”
“Because I believe it can be useful”
“The application is really interesting”
“Because I found the application easy to understand”
“Because I figured it would just make things a whole lot easier”
“I believe it would be of much help”
“It was easy to understand and operate”
“The system is easy to understand and communication is clear”
“Is quite easy and understandable”
“Because it is more easy and able to understand by a novice”
“It is easy to use”
“It is easy to use and gives some symptoms about health information”
“Because it was a bit easy to navigate around it”
“It provided a first aid guidance to minor health issues”
“It provides health information with speed and ease”.

7.7 Conclusion

From the results presented above, it can be demonstrated that the users were both satisfied with the dialogue system and that they were readily inclined to access health information using this kind of telephony system. From their responses to

the question on *why* they would like to access health information using dialogue systems, it is obvious that simplicity of use; understandability of the system [even though a white female voice was used in communicating with the participants]; usefulness of such systems; ease of use and navigation of the system; usefulness in providing first aid; and usefulness in making life easier were factors that contributed to the success of the dialogue system. Therefore, any system intended to provide health information should take into consideration these useful and practical features. This kind of system can be used to provide both cost-effective and readily available access to health information, which has heretofore been available on the Internet in text form only. This way, the Internet can better serve the populations of low literate, computer illiterate and the visually impaired, who without this Health Dialogue System would not have been able to access online information about debilitating and, in some instances, life threatening infectious diseases commonly found in Nigeria. Such systems provide a more equitable access to web-based health information for those who cannot readily access this information on line.

Acknowledgment

The author appreciates the efforts of Ogundaini Michael of the Department of Physical Sciences, Computer Science program of Landmark University, Omu-Aran, Kwara State, Nigeria in helping to administer the tests and the questionnaire.

References

- Acho, O. (2005) 'Poor healthcare system: Nigeria's moral difference', Available at: http://www.kwenu.com/publications/orabuchi/poor_healthcare.htm [Accessed 28 April 2008].
- Alan, G. B. (2004) *Elementary Statistics: A Step by Step Approach*. New York: McGrawHill, pp. 340–342.
- Alejandro, R. J. & Gagliardi A. (1998) 'Rating health information on the internet navigating to knowledge or to babel?', *JAMA*, 279:611–614.
- AlGhamdi, K. M. & Moussa, N. A. (2012) 'Internet use by the public to search for health-related information', *Int J Med Inform*, 81:363–373.
- Baker, L., Wagner, T. H., Singer, S. & Bundorf, M. K. (2003) 'Use of the Internet and e-mail for health care information: results from a national survey', *JAMA*, 289(18):2400–2406.

- Bernhardt, J. M. (2000) 'Health education and the digital divide: building bridges and filling chasms', *Health Educ Res*, 15(5):527–531.
- Bessell, T. L., McDonald, S., Silagy, C. A., Anderson, J. N., Hiller, J. E. & Sansom, L. N. (2002) 'Do Internet interventions for consumers cause more harm than good? A systematic review', *Health Expect*, 5(1):28–37.
- Bickmore, T. & Giorgino, T. (2006) 'Health dialog systems for patients and consumers', *J Biomed Inform*, 39(5):556–572.
- David, E. T. (2006) 'Press 1 to promote health behaviour with interactive voice response', *Am J Manag Care*, 12(6):305.
- Diaz, J. A., Griffith, R. A., Ng, J. J., Reinert, S. E., Friedmann, P. D. & Moulton, A. W. (2002) 'Patients' use of the Internet for medical information', *J Gen Intern Med*, 17:180–185.
- Eshareana, E. A. & Emperor, K. (2010) 'Application of ICTs in Nigerian secondary schools', *Library Philosophy Practice (e-journal)*, 1–8.
- Hadlaczky, G., Carli, V., Sarchiapone, M., Värnik, A., Balázs, J., Germanavicius, A., Hamilton, R., Wasserman, D. & Masip, C. (2013) 'Suicide prevention through internet based mental health promotion: the supreme project', *European Psychiatry*, 28(1).
- ITU (2009) *The World in 2009: ICT Facts and Figures*. Available at: http://www.itu.int/ITU-D/ict/material/Telecom09_flyer.pdf. [Accessed 20 March 2011].
- Jegede, P. O. & Owolabi, J. A. (2003) 'Computer Education in Nigerian Secondary Schools: Gaps Between Policy and Practice', *Meridian*, 6(2):1–5. Available at: <http://www.ncsu.edu/meridian/sum2003/nigeria/index.html> [Accessed 29 January 2014]
- Jesaimini, A., Rollason, V., Cedrahi, C., Luthy, C., Besson, M., Boyer, C., Desmeules, J. A. & Piguat, V. (2013) 'Searching for Health and Medication Information on the Internet. A review of the literature', *Clin Ther*, 35(8S):e17.
- Jo, H. S., Hwang, M. S. & Lee, H. (2010) 'Market segmentation of health information use on the Internet in Korea', *Int J Med Inform*, 79:707–715.
- José, R. (2007) 'Web services and speech-based applications around VoiceXML', *J Netw*, 2(1):27–35.
- Kwan, M. L. & Jennifer, L. (2005) 'Speech versus touch: a comparative study of the use of speech and DTMF keypad for navigation', *Int J Hum-Comp Int*, 19(3):343–360.
- Lintonen, T. P., Konu, A. I. & Seedhouse, D. (2008) 'Information technology in health promotion', *Health Educ Res*, 23(3):560–566.
- Lupiáñez-Villanueva, F. (2011) 'Health and the Internet: Beyond the Quality of Information', *Rev Esp Cardiol*, 64(10):849–850.
- Oyelami, O., Okuboyejo, S. & Ebiye, V. (2013) 'Awareness and usage of Internet-based health information for self-care in lagos state, Nigeria: implications for healthcare improvement', *J Health Inform Develop Count*, 7(2):165–177.
- Sadasivam, R. S., Kinney, R. L., Lemon, S. C., Shimada, S. L., Allison, J. J. & Houston, T. K. (2013) 'Internet health information seeking is a team sport: Analysis of the Pew Internet Survey', *Int J Med Inform*, 82:193–200.
- Sauro, J. & Kindlund, E. A. (2005) 'Method to Standardize Usability Metrics into a Single Score', ACM, CHI. Portland, Oregon, USA, 2–7 April.
- Walker, M., Litman, D. & Kamm, C. (1999) 'Evaluating spoken language systems', *Am Voice Input/Output Society (AVIOS)*, 25.

- Weaver III, J. B., Mays, D., Lindner, G., Eroğlu, D., Fridinger, F. & Bernhardt, J. M. (2009) 'Profiling characteristics of Internet medical information users', *J Am Med Inform Assoc*, 16(5):714–722.
- W3C. (2001) Voice Extensible Markup Language (VoiceXML) Version 2.0. Available at: <http://www.w3.org/TR/2004/REC-voicexml20-20040316/>. [Accessed 29 October 2013].
- World Health Organization. (2010) World Health Statistics 2010. Available at: http://www.who.int/gho/publications/world_health_statistics/EN_WHS10_Full.pdf [Downloaded: 23 September, 2013].

Keith M. Williams

8 DVX – the descriptive video exchange project: using crowd-based audio clips to improve online video access for the blind and the visually impaired

Abstract: In recent years, we have witnessed an explosion in the amount of online information available in video format. Full participation in an information society now requires the ability to access and understand video data. This requirement presents a major obstacle to people who can only see poorly, or not at all. As a result, their access to important medical information may be severely compromised. Furthermore, video data cannot be processed by current text-based search techniques. This chapter examines the use of audio and text description, created through crowd sourcing, to improve video accessibility for the blind and the visually impaired. In addition, it describes how description and speech recognition can improve video search.

8.1 Current problems with video data

The first problem is the difficulty of accessing video formatted data for the blind or the visually impaired. Given that more and more data are being presented in the video medium, participation in an information society requires the ability to view and understand ever-increasing amounts of video data. Blind people and the visually impaired want and need access to online video information for the same reasons as sighted people want access to such information. Here are a few examples:

Health care: An ever-increasing amount of health and medical information on the Internet is now available in video format. Nearly everything from lifestyle suggestions to prescription instructions is presented as small movies. In fact, access to this information can prove vital to one's well-being.

Education: Course materials and lectures frequently contain large amounts of information in video format. Consequently, access to video has become an educational necessity.

Entertainment: blind people want to enjoy movies, DVDs, etc.

Unfortunately, video content presents a large portion of its information visually. While highly effective, it excludes those who cannot see it. How can video format be augmented to provide increased accessibility?

The second problem is the difficulty of searching information that is in video format. Consider a patient who has a set of video instructions for a medical prescription that they must take. They want to look up the dosage. Unlike with text, the patient has no *search* function available to them that can be used to find the dosage section. Instead, the patient must tediously play through the entire video, seeking this information more or less by trial-and-error until they find the dosage information. No doubt, this is a slow, cumbersome process, whether the searcher is sighted or blind. The question posed here is how can this be made easier for everyone regardless of their visual capacity?

8.2 The description solution

Given that video is so pervasive a medium in this society, what can we do to increase its accessibility to the blind and the visually impaired, as well as improve its ability to be searched by the general population writ large? One solution is to add more information in another format to augment the video content. This is called “description.”

8.2.1 What is description?

In its broadest sense, a description is a set of information in a secondary format that augments other information that exists in a primary format.

A common example is television captioning. The television program is a stream of information in video, which is the *primary* format. The captioning is the description that adds information as text, which is the *secondary* format.

By adding a secondary format, description provides access to information when the primary format is not usable. For example, the text captioning provides information to people who are deaf, or where the environment is too noisy to use audio, as in a crowded bar.

8.2.2 Description for the visually impaired

The main objective of the DVX project, described below, is to increase accessibility to video data for the visually impaired. In the DVX project, the primary format

is video and the secondary format is audio, the addition of which provides increased access to videos for people with low or no vision.

8.2.2.1 Current types

Amateur live audio description (bring a friend)

This was, historically, the only type of description available. A blind person found a friend to watch a video with them, and the friend described the video in real time as they sat at the side of their unsighted friend.

Professional audio description

Occasionally, an organization (e.g., a film studio) that created a video would create a description to accompany it. The description became an addition to the professional product. Such descriptions were created in the same manner as other video content was created. As such, a scriptwriter would write what was to be spoken. Then, the resulting script would go through many cycles of editing and revision, and, finally, a professional speaker (voice talent) would record the script.

Problems with the current types

Description has three major technical issues, which must be addressed to work successfully: storage, distribution, and synchronization. A description's data must be stored somewhere, so it can be used more than once. Wherever the description data is stored, it must be available on demand to the system that plays the video. Thus, it must be available for distribution.

Finally, description must be synchronized to the video. Descriptions are composed of clips – short pieces of speech, which are spoken at specific times within the video. This timing is critical. Deviations of even tens of milliseconds can cause clips to interfere with the sound track of the video, resulting in reduced comprehension of both the video and the description. Larger deviations can disrupt the synchronization between the description and the events that are occurring in the video, causing a great deal of confusion.

Amateur description has serious problems with storage and distribution. The description has no storage, and must be repeated every time the video is played. The obvious solution to this problem is to record the description on a separate device, such as a digital recorder. This method fails, because it offers no mechanism to provide the necessary synchronization. Even if that problem were solved,

amateur description still lacks a means to distribute the description to every person who views the video. This raises a lot of questions: Should a DVD with the audio be available to order along with the DVD of the video itself? What about online videos? Must the description be added to the web site? In every case, the entity that makes the video available must also make the description available, but can that be practically done?

Professional description, on the other hand, usually solves the storage and distribution problems. Since the describers usually work with the content producers, the descriptions are stored and distributed with the content itself. Moreover, this approach also simplifies synchronization.

The major problems with professional description though are high cost and difficulty of scaling. Their production method creates a very refined and elegant product, but the process simply takes much too long. The cost and availability of the chain of people needed to produce professional description makes it ostensibly non-viable for small video productions. For example, an academic department at a university setting that wanted to describe videos of its seminars and lectures would no doubt lack the time and money to hire competent scriptwriters.

8.3 Architecture of DVX

The objective of the DVX project was to provide an environment, or platform, for the creation and dissemination of video description, which addressed the problems stated above, and more. To achieve this, the architecture is divided into two distinct parts. The first part is the Descriptive Video Exchange (DVX) server, which provides an open, distributed repository for descriptions. The second part includes the Applications Modules that utilize the DVX server to provide description solutions for their users.

8.3.1 The DVX server

The DVX server is the common repository that stores and disseminates video descriptions. It exposes a standard interface that applications can use to store, manage, and distribute description data. It does this in the form of a RESTful web service, which is an architectural style consisting of a coordinated set of architectural constraints applied to components, connectors, and data elements, within a distributed system.

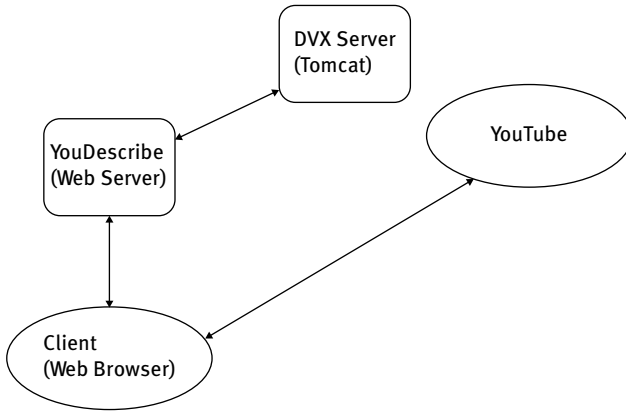


Fig. 8.1: DVX architecture.

8.3.1.1 Major data elements, attributes and actions

Figure 8.1 shows the three major data elements in the server: user, clip, and video. Each is specified by a set of attributes shown in Tab. 8.1. Together, they encapsulate the information and inter-relationships that clients of the server need in order to create and manage descriptions. Note: each element has more attributes than are listed here. These are just some of the major ones.

Tab. 8.1: DVX data elements.

Element	Attributes	Actions
User	Handle – the user’s name Password – the user’ identification Email – user’s communication point	Add – create a new user Logon – sign onto the server Logoff – sign off of the server
Clip	Id – a unique identifier Format – the type of the clip Video – identifies the video the clip belongs to Time – when the clip should be played Filename – the file containing the clip’s data Author – the clip’s creator	Upload – add a clip Download – retrieve a clip Metadata – provide information on a clip Delete – remove a clip
Video	Id – identifies the video Media identifier – identifies the exact version of the video Author – the user who added a video	Add – enters a new video into the server Video – retrieves information about a video

User

The DVX server supports the concept of a user. Users represent the people who create descriptions. Each user has a unique identifier, as well as a name and password. Users must log in to the server to make any changes to data (add, delete, modify). They do not have to log in to simply query or view data. An “id” number uniquely identifies each user within the server.

Attributes

A user has a “handle,” which will be the name he or she is known by to the server.

A user’s “password” provides verification of user identity.

A user also has an “email” address.

Actions

A client can “add” a user to the DVX server. This starts a registration process that sends a confirmation message to the given email address so as to confirm the identity of the candidate user.

A client can “logon” to the server, using the “handle” and “password” attributes.

A client can “logoff” from the server.

Clip

A clip is a piece of information that says something about a video. A set of clips, each played at a different time in the same video, forms a description. The server allows clips to be in many different formats: audio files, text strings, etc.

Attributes

Each clip has several parameters associated with it.

A clip’s “id” uniquely identifies it.

Its “format” indicates what the medium of the clip is: audio, text, etc.

The “video” attribute is the id of the video that the clip belongs to.

The “time” of a clip defines the time, in seconds, from the beginning of the video that the clip belongs to, i.e., the time within the video that clip should be played.

The “filename” is the name of the file that contains the clip’s data.

Finally, the “author” is the id of the user who created the clip. Each clip belongs to a particular user. This allows multiple users to create separate descriptions of the same video.

Actions

Clients of the DVX server can perform four actions with clips.

Given the attributes “filename,” “time,” “author” and “video,” the “upload” action copies “filename” into the server.

The “download” action streams the contents of a clip to the requesting client. Any combination of attributes can be specified, as long as they define a unique clip.

A client can request a “Metadata” query for clips. This returns information about *all* clips (as opposed to the clips themselves) that match the parameters passed with the query. If no parameters are passed, information on every clip in the server is returned.

The “delete” action removes a clip from the server. As with download, any number of attributes can be specified, as long as they identify a unique clip.

Video

The video element contains information about a video. It does not contain the video itself.¹

Attributes

The “id” attribute is an internal identifier for each video known to the server. The “media identifier” is a unique name for each video. As opposed to the “id” mentioned above, this is an id, found from the media or its source itself that identifies it as a particular piece of content. For example, it identifies a video as “great video, version 1.” This is important, because content with the same title, etc. can often be found in different versions. This is particularly common with DVDs – there is an original version, a “director’s cut,” a PG-edited version, etc. These versions will almost surely have different lengths.

In order to maintain synchronization, it is vital that the server knows the exact identity of each piece of content that is described. Websites, such as YouTube, often have unique identifiers for each video, making the solution relatively easy.

¹ Notice that these actions lack a method to store the video within server, and provide no mechanism to alter videos in any way. This was a conscious design decision. The DVX architecture enables people to create descriptions, who are not the creators or owners of the content they describe. For this reason, DVX protects the content from any possible changes. The DVX server does not have any capability to download, store or alter the described media. It only stores and manipulates descriptions, and merely points, through URLs, to the described content. All content remains untouched at their original locations.

Other media, such as DVDs, can be more difficult. There are sometimes tags within the DVD that are meant to be unique, but these are not always present or accurate. We have explored the use of hashing algorithms to derive unique keys from the actual data stream from DVDs. This is an ongoing area of research. The “author” attribute is the id of the user who added the video to the server. This is separate from the people or group who created the video.

Actions

The “add” action lets a client add a new video to the server.

The “video” action retrieves information about all videos that match the given attributes.

8.3.1.2 Current implementation

RESTful web service interface

A web service can be described as a web site whose clients are other computers, rather than humans. Programs communicate with the site to access and exchange data, as well as to perform actions on that data. A RESTful web service follows the paradigm given by Roy Fielding² in 2000. Programs communicate with the service using the common HTTP protocol.

The server exposes URLs to clients over a network. In each URL, the action to be taken is presented as a combination of an HTTP method and an endpoint, and the attributes are represented by HTTP parameters. For example, a client request for a particular clip would have the form: `http://dvxwebsite.com:8080/dvxApi/clipdownload?video=1234&Time=32.6&Author=25`.

This is an HTTP GET method that, which instructs the DVX server (hosted at `dvxwebsite.com`), to download the clip created by user number 25, for the video with an internal identification of 1234, which was recorded 32.6 seconds from the start of the video.

8.3.1.3 Tomcat servlet container

The DVX server resides in a Tomcat container. Tomcat is an open-source platform similar to a web server such as Apache or Windows IIS. Instead of hosting HTML web pages, however, it hosts Java objects called servlets. These objects respond to HTTP requests, and dynamically create HTTP responses. The DVX

² Fielding, Roy Thomas (2000), *Architectural Styles and the Design of Network-based Software Architectures*, Doctoral dissertation, University of California, Irvine.

server is a collection of servlets that receive HTTP requests from clients, and reply with responses.

MySQL database/Hibernate

All data except for the actual clips are stored on a MySQL database. MySQL provides the persistent storage needed to preserve the relationships between clips and videos that comprise descriptions, as well as the information required to manage users, videos and other bookkeeping activities. Hibernate is an Object Relational Mapping (ORM) framework. It facilitates the translation between the object-oriented environment of the Java code in Tomcat, and the relational data paradigm of MySQL.

8.3.1.4 Applications

The DVX server provides vital description information programmatically, in a form that is easy for software programs to work with. The second half of the DVX architecture is its applications. These are the software systems that utilize the DVX server to provide description capabilities and functions to users.

YouDescribe

YouDescribe is a description creation and distribution web application for YouTube videos. Anyone with a web browser can use it.

A user can select any normal YouTube video, and play it. At any point, they can pause the video and record an audio clip. By repeating this process, they create a description of the video.

Later, another user can select and play the same video. As it plays, YouDescribe plays the audio clips for them at the appropriate times in the video.

Figure 8.2 shows the main YouDescribe page.

If the user clicks on the “Show Most Recent YouDescribe Videos” link, a table displays all the YouTube videos that have descriptions. The table shows a row for each description of each video; therefore some videos are displayed on multiple rows.

Additionally, the user can enter text into the search textbox. The table will then show all YouTube videos that match the search criteria.

In either case, each row of the table displays an icon of the video, followed by its title. Clicking on either of these pops up a video player, which starts to play the selected video (Fig. 8.3).

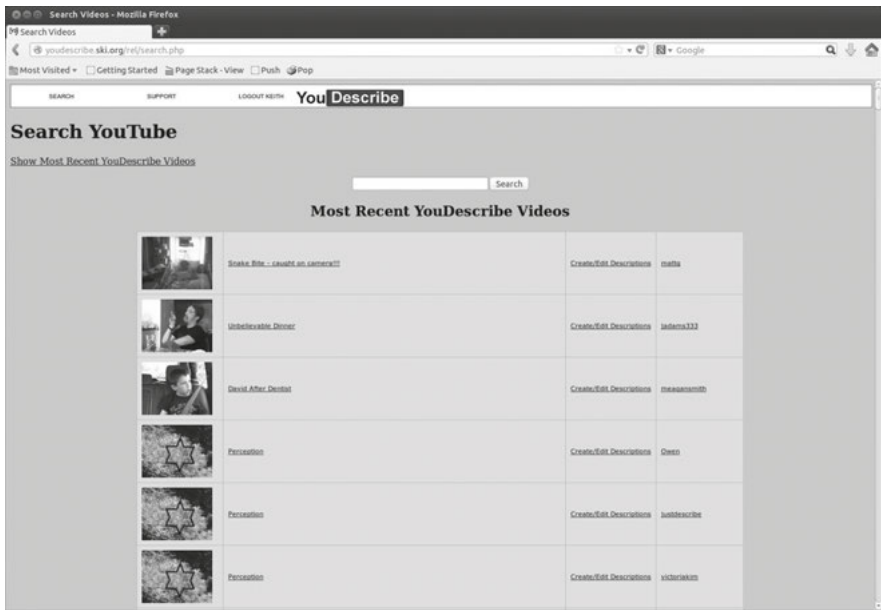


Fig. 8.2: YouDescribe.



Fig. 8.3: YouDescribe player.

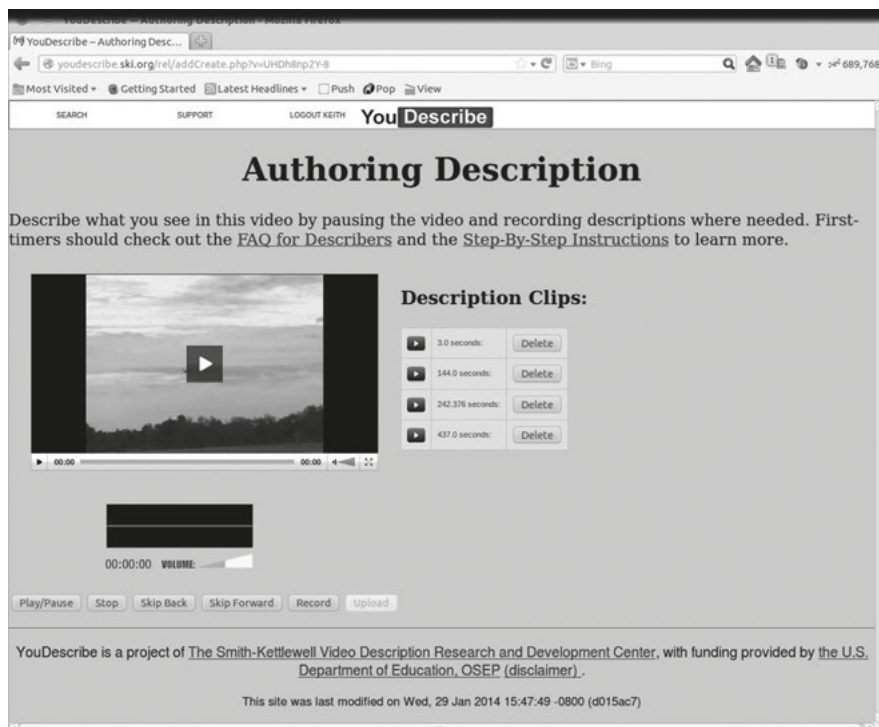


Fig. 8.4: YouDescribe authoring.

The player floats on top of the main page. It behaves similarly to any ordinary video player, except that while playing the video, it will pause and play audio clips at appropriate times, which provides a description of the video. The player also includes a describer dropdown, and a “share” button. The dropdown allows the user to select the describer they want to hear, while the “share” button displays a link to the described video, which a user can use to embed the video in any web document.

Returning to the main page, an optional third field is only displayed if the user is logged in to the system. Selecting it brings up the authoring page (Fig. 8.4), where the user can record descriptions for the selected video.

The authoring page allows a logged-in user to create and edit a video’s description, by recording and deleting individual clips. It displays a video player, similar to the one in the main page, except for four additional controls:

- A record button, when pressed, will record the user’s speech.

- A volume control allows the user to adjust the recording level.

A small oscilloscope displays shows a waveform of what is being recorded. An upload button, when pressed, will send the recording to YouDescribe.

To author a description, a user would play the video, and pause it at a point where they want to describe something. They would then hit the record button, speak their explanation of the video, and click the upload button to send the clip to YouDescribe. YouDescribe then sends the clip to DVX, which stores it as a clip for the video, to be played back in the future at the time where the video is paused.

An additional frame in the authoring page lists all the clips the user has created for the video. Each line of the list allows the user to play back the clip, and delete it, if desired.

The final field in the main page table shows the name of the user who created descriptions for the video. Clicking on that name returns a table, which lists all the videos that user described. This provides a way to follow a particular user, and track the videos they have described.

An example scenario – Figure 8.5

An example will help illustrate how YouDescribe can use the DVX server to provide description services.

When a user goes to the YouDescribe web page, the browser sends a request to the DVX server, in the form:

HTTP method: GET
Endpoint: Video

This retrieves information about every video known to YouDescribe. The web page displays the returned information as a list. There is an entry in the list for each video that has a description created by a particular user. Therefore, a video may be listed several times, if more than one user created a description for it.

The user selects a video from the list. The browser creates a player to play the video. It then queries the server, to get a list of all the clips associated with that video:

HTTP method: GET
Endpoint: clip/metadata
Attributes: Id=“id of selected video,” Author=“id of selected user”

The server responds with data for every clip that was recorded by that user for that video. The browser then starts the player, and plays the video.

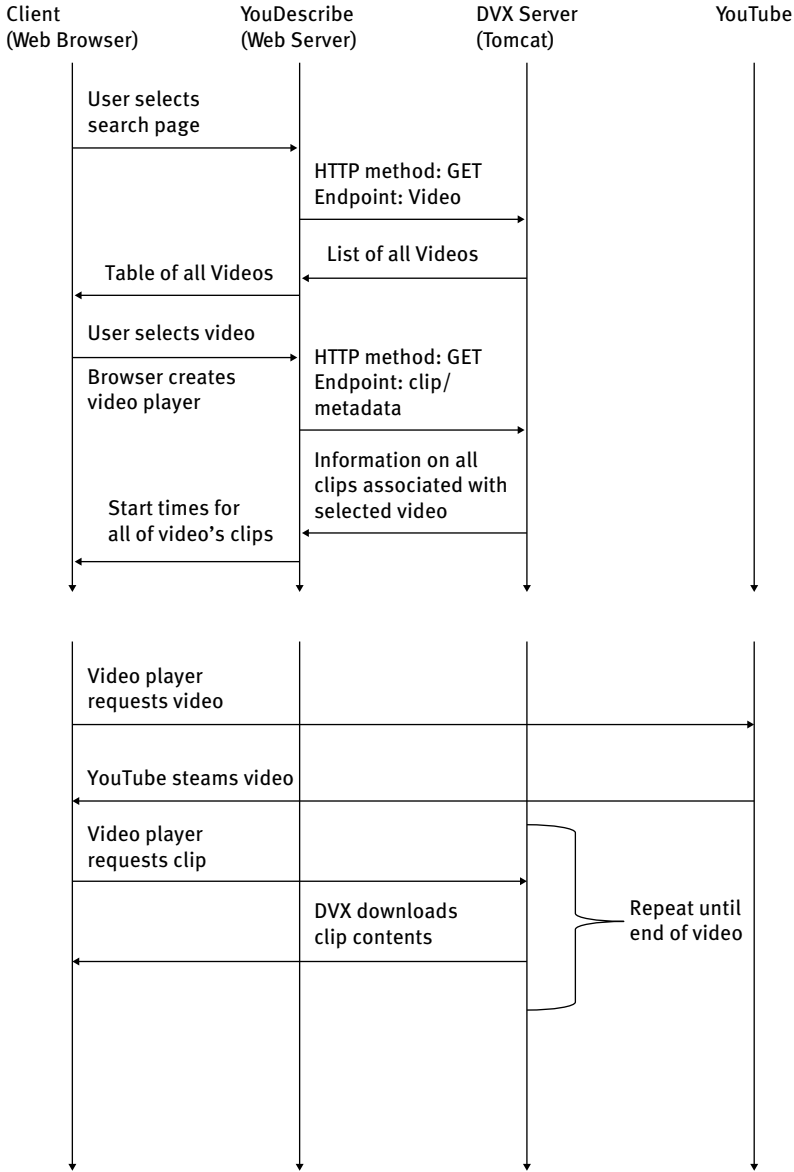


Fig. 8.5: Example scenario.

Because the data returned by the server includes a “time” attribute for each clip, the player knows when it should pause the video and play an audio clip. Whenever it is time to play a clip, the browser requests the clip from the server:

HTTP method: GET

Endpoint: clip

Attributes: Id=“id of selected video,” Author=“id of selected user,”

Time=“time of desired clip”

The browser then pauses the player, and plays the clip, which describes what occurs in the video at that time. When the clip is done, the browser tells the player to resume playing. In this manner, the application provides a description of the video.

8.4 DVX solves description problems

The combination of the DVX server and DVX-enabled applications solves many of the problems associated with description creation and dissemination:

First, it allows anyone to create descriptions. The DVX Server, combined with an application such as YouDescribe, creates a crowd-sourced solution for description generation. It grants the ability to describe video to every person with a web browser. This includes professional describers. The DVX interface augments the tools professional describers already use, thereby providing a gateway to DVX’s instant, worldwide distribution and storage.

Second, it provides description distribution. DVX breaks the traditional tie between description creation and a particular vendor or video medium. The origin of a description is independent of its storage and access. The descriptions are stored where any user with Internet connectivity can access them.

Third, it provides synchronization information. In DVX, all clips are stored with the time that they should be played in a video. Therefore, correct time information is always available to any application, which eases the task of description synchronization.

Fourth, it enforces common description formats. In DVX, all clips are stored in a common format. Descriptions can be easily shared regardless of the different applications that generated them.

Fifth, it decreases the time it takes to distribute descriptions. A description, once created, is immediately available. It does not have to be delayed, pending a new release of the content.

8.5 DVX and video search

While DVX can help solve many issues concerning video description, it can also be used to improve video search. As mentioned before, searching a video is a very difficult process. Imagine, however that the video had descriptions in text format, commonly known as “tags.” Those descriptions are available to the many text search tools on the market today. A search for the word “dosage” could quickly forward the video to the information one needs with regard to how much and how often an especially potent medication may be taken. Further, data mining tools, in addition to all the other powerful analytical techniques that have been developed for text data, can now process the video data as well.

The challenge is how to create tags for the vast number of videos.

Speech recognition has been used to create a text transcript of a video’s audio track. Search engines and other text-based tools can then access and process the transcript. This approach is fast, since it requires little or no human intervention, and it can work nearly automatically.

However, automatically creating tags has two shortcomings:

First, the transcript only contains information from the audio portion of the video. Thus it may omit vital information that could have been used to create useful indexing tags. For example, the prescription video may never actually say the word “dosage.” Just as with a visually-impaired viewer, the speech recognition system (which cannot see either) only knows what is contained on the audio track. Second, the transcription is not constituted or formulated as a set of tags; it is merely a text translation of all the voice on the video. Consequently, it must be processed further to extract a useful set of tags.

DVX could bring its solutions to greatly enhance the transcript approach. First, it could fill the gap that the transcript misses. The descriptions people create are information about the visual-only portions of the video, which is the very part the transcript doesn’t cover. It could use speech recognition on those descriptions to create a second set of text that fills in the information that is missing from the transcription. Second, the descriptions it translated to text were created by humans, who used their judgment to create intelligent translations of voice text, thereby making it much easier to extract useful tags than to rely on machine learning techniques to make these feature extractions.

Futhermore, DVX could also create tags directly. An application could be written that simply enables users to pause a video and type a description, rather than speak one. The DVX server can store any type of clip, as long as it can be stored as a file. DVX could even convert those descriptions to audio, using text-to-speech.

8.6 Conclusion

Description is a powerful tool that increases accessibility to information that is stored in video format. The Descriptive Video Exchange provides a framework that enables a large number of people, amateur and professional, to create descriptions both quickly and easily. It distributes those descriptions so that they are available to anyone on the Internet and, in particular, provides a special service for the visually impaired. Furthermore, DVX when combined with speech recognition can greatly improve video search.

Acknowledgment

The project described in this chapter was developed by the Smith-Kettlewell Video Description Research and Development Center, under a grant from the U.S. Department of Education (H327J110005), awarded to Dr. Joshua Miele, Principal Investigator at Smith-Kettlewell. However, these contents do not necessarily represent the policy of the U.S. Department of Education and you should not assume endorsement by the Federal Government.

**Part IV Visual data: new methods and approaches
to mining radiographic image data and
video metadata**

9 Information extraction from medical images: evaluating a novel automatic image annotation system using semantic-based visual information retrieval

Abstract: Today, in the medical field there are huge amounts of non-textual information, such as radiographic images, generated on a daily basis. Given the substantial increase of medical data stored in digital libraries, it is becoming more and more difficult to perform search and information retrieval tasks. Image annotation remains a difficult task for two reasons: (1) the semantic gap problem – it is hard to extract semantically meaningful entities when using low-level image features; and (2) the lack of correspondence between the keywords and image regions in the training data. Content-based visual information retrieval (CBVIR) and image annotation has attracted a lot of interest, namely from the image engineering, computer vision, and database community. Unfortunately, current methods of the CBVIR systems only focus on appearance-based similarity, i.e., the appearance of the retrieved images is similar to that of a query image. As a result, there is very little semantic information exploited. To develop a semantic-based visual information retrieval (SBVIR) system two steps are required: (1) to extract the visual objects from images; and (2) to associate semantic information with each visual object. The first step can be achieved by using segmentation methods applied to images, while the second step can be achieved by using semantic annotation methods applied to the visual objects extracted from images. In this chapter, we use original graph-based color segmentation methods because we find that as linear algorithms they perform well. The annotation process implemented in our system is based on the Cross-Media Relevance Model (CMRM), which invokes principles defined for relevance models. For testing the annotation module, we have used a set of 2000 medical images: 1500 of images in the training set and 500 test images. For testing the quality of our segmentation algorithm, the experiments were conducted using a database consisting of 500 medical images of the digestive system that were captured by an endoscope. Our initial test results, based on looking at the assigned words to see if they were relevant to the image in question, have proven that our automatic image

annotation system augurs well in the diagnostic and treatment process. This is first step toward larger studies of automatic image annotation for indexing, retrieving, and understanding large collections of image data.

9.1 Introduction

Advances in medical technology generate huge amounts of non-text information (e.g., images) along with more familiar textual one. Medical images play a central role in patient diagnosis, therapy, surgical planning, medical reference, and medical training. The image is one of the most important tools in medicine since it provides a method for diagnosis and monitoring of patients' illnesses and conditions, with the advantage of it being a very fast, non-invasive procedure. *Automatic image-annotation* (also known as automatic image tagging or linguistic indexing) is the process by which the computer system automatically assigns metadata, in the form of captioning or keywords, to the digital image while taking into account its content. This process is of great value as it allows indexing, retrieving, and understanding of large collections of image data. As new image acquisition devices are continually developed to produce more accurate information and increase efficiency, and as data storage capacity likewise increases, a steady growth in the number of medical images produced can be easily inferred. Given the massive increase of medical data in digital libraries, it is becoming more and more difficult to perform search and information retrieval tasks. In sum, image annotation remains a difficult task for two main reasons: (1) the semantic gap problem – it is hard to extract semantically meaningful entities when using low-level image features; and (2) the lack of correspondence between the keywords and image regions in the training data.

Recently, there has been lot of discussion about semantically-enriched information systems, especially about using ontology for modeling data. In this chapter, we present a novel image-annotation system, revolving on a more comprehensive information extraction approach, for use in the medical domain. The annotation model used was inspired from the principles defined for the cross-media relevance model. The ontology used by the annotation process was created in an original manner starting from the information content provided by the medical subject headings (MeSH). Our novel approach is based on the double assumption that given images from digestive diseases, expressing all the desired features using domain knowledge is feasible; and that manually marking up and annotating the regions of interest is practical as well. In addition, by developing an automatic annotation system, representing and reasoning about medical

images are performed with reasonable complexity within a given query context. Not surprisingly, due to the presence of a large number of images without text information, content-based medical image retrieval has received quite a bit of attention in recent years.

9.2 Background

Content-based visual information retrieval (CBVIR) has attracted a lot of interest, namely from the image engineering, computer vision and database community. A large corpus of research has been built up in this field showing substantial results. Content-based image retrieval task could be described as a process for efficiently retrieving images from a collection by similarity. The retrieval relies on extracting the appropriate characteristic quantities describing the desired contents of images. Most CBVIR approaches rely on the low-level visual features of image and video, such as color, texture and shape. Such techniques are called *feature-based techniques* in visual information retrieval (Tousch et al. 2012). Unfortunately, current methods of the CBVIR systems only focus on appearance-based similarity, i.e., the appearance of the retrieved images is similar to that of a query image. As a result, there is very little semantic information exploited. Among the few efforts which claim to exploit the semantic information, the semantic similarities are defined between different appearances of the same object. These kinds of semantic similarities represent the low-level semantic similarities, while the similarities between different objects represent the high-level semantic similarities. The similarities between two images are the similarities between the objects contained within the two images. As a consequence, a way to develop a semantic-based visual information retrieval (SBVIR) system consists of two steps: (1) to extract the visual objects from images; and (2) to associate semantic information with each visual object. The first step can be achieved by using segmentation methods applied to images, while the second step can be achieved by using semantic annotation methods applied to the visual objects extracted from images.

Image segmentation techniques can be separated into two groups: region-based and contour-based approaches. Region-based segmentation methods can be broadly classified as either top-down (model-based) or bottom-up (visual feature-based) approaches (Adamek et al. 2005).

An important group of visual feature-based methods is represented by the graph-based segmentation methods, which attempt to search for certain structures in the associated edge-weighted graph constructed on the image pixels, such as

minimum spanning tree or minimum cut. Other approaches to image segmentation which are region-based consist of splitting and merging regions according to how well each region fulfills some uniformity criterion. Such methods use a measure of uniformity of a region. In contrast, other region-based methods use a pair-wise region comparison rather than applying a uniformity criterion to each individual region. Liew & Yan (2005) demonstrate that the contour-based segmentation approach, as distinguished from region-based approach, assumes that different objects in an image can be segmented by detecting their boundaries. The authors further sharpen the distinction between these two groups: “whereas region-based techniques attempt to capitalize on homogeneity properties within regions in an image, boundary-based techniques [used in the contour-based approach] rely on the gradient features near an object boundary as a guide” (p. 316).

Medical images segmentation describes some graph-based color segmentation methods, and an area-based evaluation framework of the performance of the segmentation algorithms.

The proposed SBVIR system involves an annotation process of the visual objects extracted from images. It becomes increasingly expensive to manually annotate medical images. Consequently, automatic medical image annotation becomes important. We consider image annotation as a special classification problem, i.e., classifying a given image into one of the predefined labels.

Several interesting techniques have been proposed in the image annotation research field. Most of these techniques define a parametric or non-parametric model to capture the relationship between image features and keywords (Stumme & Maedche 2001). The concepts used for annotation of visual objects are generally structured in hierarchies of concepts that form different ontologies. The notion of ontology is defined as an explicit specification of some conceptualization, while the conceptualization is defined as a semantic structure that encodes the rules of constraining the structure of a part of reality. The goal of ontology is to define some primitives and their associated semantics in some specified context. Ontology has been established for knowledge sharing and is widely used as a means for conceptually structuring domains of interest. With the growing usage of ontology, the problem of overlapping knowledge in a common domain occurs more often and becomes critical. Domain-specific ontology is modeled by multiple authors in multiple settings. Such an ontology lays the foundation for building new domain specific ontology in similar domains by assembling and extending ontology from repositories. Though ontology is frequently used in the medical domain, existing ontology is provided in formats that are not always easy to interpret and use. To handle these uncertainties, researchers have proposed

a great number of annotation models and information extraction techniques (Stanescu et al. 2011).

9.3 Related work

Because ontology are not always easy to interpret, a number of models using a discrete image vocabulary have been proposed for image annotation (Mori, Takahashi & Oka 1999; Duygulu et al. 2002; Barnard et al. 2003; Blei & Jordan 2003; Jeon, Lavrenko & Manmatha 2006; Lavrenko, Manmatha & Jeon 2006). One approach to automatically annotating images is to look at the probability of associating words with image regions. Mori, Takahashi & Oka (1999) used a co-occurrence model where they looked at the co-occurrence of words with image regions created using a regular grid. To estimate the correct probability this model required large numbers of training samples. Thus, the co-occurrence model, translation model (Duygulu et al. 2002), and the cross-media relevance model (CMRM) (Jeon, Lavrenko & Manmatha 2003) demonstrates that each is respectively trying to improve a previous model.

Annotation of medical images requires a nomenclature of specific terms retrieved from ontology to describe its content. For medical domain what can be used is either an existing ontology named open biological and biomedical ontology (<http://www.obofoundry.org/>) or a customized ontology based on a source of information from a specific domain.

The medical headings (MeSH) (<http://www.nlm.nih.gov/>) and (http://en.wikipedia.org/wiki/Medical_Subject_Headings) are produced by the National Library of Medicine (NLM) and contain a high number of subject headings, also known as descriptors. MeSH thesaurus is a vocabulary used for subject indexing and searching of journal articles in MEDLINE/PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>). MeSH has a hierarchical structure (http://www.nlm.nih.gov/mesh/2010/mesh_browser/MeSHtree.html) and contains several top level categories like anatomy, diseases, health care, etc. Relationships among concepts (<http://www.nlm.nih.gov/mesh/meshrels.html>) can be represented explicitly in the thesaurus as relationships within the descriptor class. Hierarchical relationships are seen as parent-child relationships and associative relationships are represented by the “see related” cross reference.

Duygulu et al. (2002) described images using a vocabulary of blobs (which are clusters of image regions obtained using the K-means algorithm). Image regions were obtained using the normalized-cuts segmentation algorithm. For each image region 33 features such as color, texture, position and shape information were

computed. The regions were clustered using the K-means clustering algorithm into 500 clusters called “blobs.” This annotation model called translation model was a substantial improvement of the co-occurrence model. It used the classical IBM statistical machine translation model (Brown et al. 1993) making a translation from the set of blobs associated to an image to the set of keywords for that image.

Jeon et al. (2003) viewed the annotation process as analogous to the cross-lingual retrieval problem and used a cross media relevance model to perform both image annotation and ranked retrieval. The experimental results have shown that the performance of this model on the same dataset was considerably better than the models proposed by Duygulu et al. (2002) and Mori et al. (1999).

There are other models like correlation LDA proposed by Blei & Jordan (2003) that extends the Latent Dirichlet Allocation model to words and images. This model is estimated using expectation-maximization algorithm and assumes that a Dirichlet distribution can be used to generate a mixture of latent factors. In Li & Wang (2003) it is described a real-time ALIPR image search engine which uses multi resolution 2D hidden Markov models to model concepts determined by a training set. In an alternative approach, Blei & Jordan (2003) rely on a hierarchical mixture representation of keyword classes, leading to a method that has a computational efficiency on complex annotation tasks. There are other annotation systems used in the medical domain like I2Cnet (Image indexing by Content network) Catherine, Xenophon & Stelios (1997) providing services for the content-based management of images in health care. In Igor et al. (2010), the authors present a hierarchical medical image annotation system using Support Vector Machines (SVM) – based approaches.

In Daniel (2003) the author provides an in depth description of Oxalis, a distributed image annotation architecture allowing the annotation of an image with diagnoses and pathologies. In Baoli, Ernest & Ashwin (2007) the authors describe the SENTIENT-MD (Semantic Annotation and Inference for Medical Knowledge Discovery) a new generation medical knowledge annotation and acquisition system.

In Peng, Long & Myers (2009) the authors present VANO, a cross-platform image annotation system enabling the visualization and the annotation of 3D volume objects including nuclei and cells.

Some of the elements (e.g., the clustering algorithm used for obtaining blobs or the probability distribution of the CMRM) included in our system are related to *Soft Computing* which is an emerging field that consists of complementary elements of fuzzy logic, neural computing, evolutionary computation, machine

learning and probabilistic reasoning. Machine learning includes unsupervised learning which models a set of inputs, like clustering.

9.4 Architecture of system

The annotation process implemented in our system is based on the cross-media relevance model (CMRM), which invokes principles defined for relevance models. Using a set of color-annotated images of the diseases of the digestive system, the system learns the distribution of the blobs and words. The diseases are indicated in images by color and texture changes. Having the set of blobs, each image from the test set is then represented using a discrete sequence of blobs identifiers. The distribution is used to generate a set of words for a new image.

The architecture of our system is presented in Fig. 9.1 and contains six modules (Burdescu et al. 2013):

- *Segmentation module* – these modules segment an image into regions by planar segmentation; it can be configured to segment all images from an existing images folder on the storage disk. The hexagonal structure used by the owner segmentation algorithm represents a grid-graph and is presented

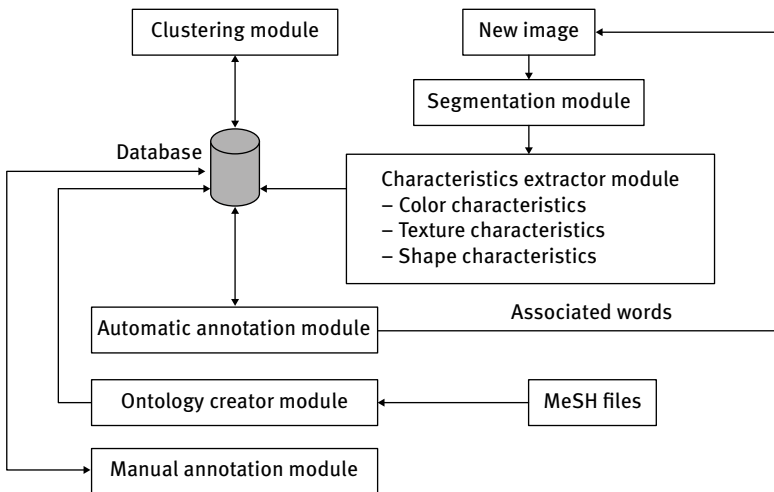


Fig. 9.1: System's architecture.

in Fig. 9.2. For each hexagon “h” in this structure there exist 6-hexagons that are neighbors in a 6-connected sense. The segmentation process is using some methods in order to obtain the list of regions:

- Same vertex color – used to determine the color of a hexagon
- Expand colour area – used to determine the list of hexagons having the color of the hexagon used as a starting point and has as running time where n is the number of hexagons from a region with the same color.
- List regions – used to obtain the list of regions and has as running time where n is the number of hexagons from the hexagonal network.
- *Characteristics extractor module* – this module is using the regions detected by the Segmentation module. For each segmented region is computed a feature vector that contains visual information of the region such as color (color histogram with 166 bins, texture (maximum probability, inverse difference moment, entropy, energy, contrast, correlation), position (minimum bounding rectangle) and shape (area, perimeter, convexity, compactness). The components of each feature vector are stored in the database.
- *Clustering module* – we used K-means with a fixed value of 80 (established during multiple tests) to quantize these feature vectors obtained from the training set and to generate blobs. After the quantization, each image in the training set is represented as a set of blobs identifiers. For each blob it is computed a median feature vector and a list of words belonging to the test images that have that blob in their representation.

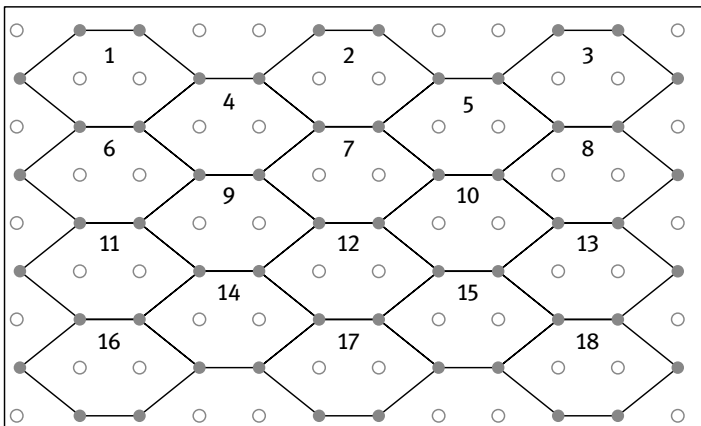


Fig. 9.2: Hexagonal structure constructed on the image pixels.

- *Annotation module* – for each region belonging to a new image it is assigned the blob which is closest to it in the cluster space. The assigned blob has the minimum value of the Euclidian distance computed between the median feature vector of that blob and the feature vector of the region. In this way the new image will be represented by a set of blobs identifiers. Having the set of blobs and for each blob having a list of words we can determine a list of potential words that can be assigned to the image. What needs to be established is which words describe better the image content. This can be made using the formulas of the cross media relevance model:

$$P(w|J) = (1 - \alpha_j) \frac{\#(w,J)}{|J|} + \alpha_j \frac{\#(w,T)}{|T|} \quad (1)$$

$$P(b|J) = (1 - \beta_j) \frac{\#(b,J)}{|J|} + \beta_j \frac{\#(b,T)}{|T|} \quad (2)$$

where:

- $P(w|J)$, $P(b|J)$ denote the probabilities of selecting the word “w,” the blob “b” from the model of the image J .
- $\#(w,J)$ denotes the actual number of times the word “w” occurs in the caption of image J .
- $\#(w,T)$ is the total number of times “w” occurs in all captions in the training set T .
- $\#(b,J)$ reflects the actual number of times some region of the image J is labeled with blob “b.”
- $\#(b,T)$ is the cumulative number of occurrences of blob “b” in the training set.
- $|J|$ stands for the count of all words and blobs occurring in image J .
- $|T|$ denotes the total size of the training set.

The smoothing parameters alpha and beta determine the degree of interpolation between the maximum likelihood estimates and the background probabilities for the words and the blobs, respectively. The values determined after experiments for the cross media relevance model were alpha = 0.1 and beta = 0.9. For each word is computed the probability to be assigned to the image an after that the set of “n” (configurable value) words having a high probability value will be used to annotate the image. We have used five words for each image.

Ontology creator module – this module has as input the MeSH content that can be obtained from (<http://www.nlm.nih.gov/mesh/filelist.html>) and is offered as an “xml” file named *desc2010.xml* (2010 version) containing the descriptors and a “txt” file named *mtrees2010.txt* containing the hierarchical structure.

This module generates the ontology and stores it in the database. This module also offers the possibility to export the ontology content as a Topic Map (<http://www.topicmaps.org/>) by generating an *.xtn file using the “xtn” syntax.

The ontology contains:

- Concepts – each descriptor is mapped to an ontology concept having as unique identifier the content of the *DescriptorUI xml* node. The name of the concept is retrieved from the *DescriptorName xml* node. The tree node of this concept in the hierarchical structure of the ontology is established using the tree identifiers existing in the *TreeNumber xml* nodes. Usually a MeSH descriptor can appear in multiple trees. For the descriptor mentioned in the above example the concept will have the following properties
 - id:D000001, name:Calcimycin, tree_nodes: D03.438.221.173
- Associations defined between concepts – our ontology contains two types of associations:
 - parent-child – generated using the hierarchical structure of the MeSH trees and the tree identifiers defined for each concept (used to identify the concepts implied in the association)
 - related-to – a descriptor can be related to other descriptors. This information is mentioned in the descriptor content by a list of *DescriptorUI* values. In practice a disease can be caused by other diseases.
- *Manual annotation module* – this module is used to obtain a training set of annotated images needed for the automatic annotation process. This module is usually used after the following steps are completed:
 - the doctor obtains a set of images collected from patients using an endoscope and this set is placed in a specific disk location that can be accessed by our segmentation module
 - the segmentation module segments each image from the training set
 - the set of regions obtained after segmentation is processed by the characteristics’ extractor module and all characteristic vectors are stored in the database
 - the clustering module using the k-means algorithm generates the set of blobs and each image is represented by a discrete set of blobs.

The manual annotation module has a graphical interface, which allows the doctor to select images from the training set to see the regions obtained after segmentation and to assign keywords from the ontology created for the selected image.

9.5 The segmentation algorithm – graph-based object detection (GBOD)

Segmentation is the process of partitioning an image into non-intersecting regions such that each region is homogeneous and the union of no two adjacent regions is homogeneous. Formally, segmentation can be defined as follows.

Let F be the set of all pixels/voxels and $P()$ be a uniformity (homogeneity) predicate defined on groups of connected pixels/voxels, then segmentation is a partitioning of the set F into a set of connected subsets or regions (S_1, S_2, \dots, S_n) such that $\cup_{i=1}^n S_i = F$ with $S_i \cap S_j = \emptyset$ when $i \neq j$. The uniformity predicate $P(S_i)$ is true for all regions S_i and $P(S_i \cup S_j)$ is false when S_i is adjacent to S_j .

This definition can be applied to all types of images.

The goal of segmentation is typically to locate certain objects of interest which may be depicted in the image. Segmentation could therefore be seen as a computer vision problem. A simple example of segmentation is to threshold a grayscale image with a fixed threshold “ t ”: each pixel/voxel “ p ” is assigned to one of two classes, P_0 or P_1 , depending on whether $I(p) < t$ or $I(p) \geq t$.

Grouping can be formulated as a graph partitioning and optimization problem by Pushmeet Kohli et al. (2012) and C. Allène et al. (2010).

The graph theoretic formulation of image segmentation is as follows:

1. The set of points in an arbitrary feature space are represented as a weighted undirected graph $G = (V, E)$, where the nodes of the graph are the points in the feature space
2. An edge is formed between every pair of nodes yielding a dense or complete graph.
3. The weight on each edge, $w(i, j)$ is a function of the similarity between nodes i and j .
4. Partition the set of vertices into disjoint sets V_1, V_2, \dots, V_k where by some measure the similarity among the vertices in a set V_i is high and, across different sets V_i, V_j is low.

To partition the graph in a meaningful manner, we also need to:

- Pick an appropriate criterion (which can be computed from the graph) to optimize, so that it produces a clear segmentation.
- Finding an efficient way to achieve the optimization.

In the image segmentation and data clustering community, there has been a substantial amount of previous work using variations of the minimal spanning tree or limited neighborhood set approaches (Grundmann et al. 2010). Although

such approaches use efficient computational methods, the segmentation criteria used in most of them are narrowly based on local properties of the graph. However, because perceptual grouping is about extracting the global impressions of a scene, this partitioning criterion often falls short of this main goal.

There are huge of papers for 2D images and segmentation methods and most graph-based for 2D images and few papers for spatial segmentation methods. Because we used an original segmentation algorithm, we depict only the set of hexagons constructed on the image Burdescu et al. (2009); Brezovan et al. (2010); and Stanescu et al. (2011).

The method we use pivots on a general-purpose segmentation algorithm, which produces good results from two different perspectives: (1) from the perspective of perceptual grouping of regions from the natural images (standard RGB); and (2) from the perspective of determining regions if the input images contain salient visual objects.

Let $V = \{h_1, \dots, h_{|V|}\}$ be the set of hexagons/tree-hexagons constructed on the spatial image pixels/voxels as presented above and $G = (V, E)$ be the undirected spatial grid-graph, with E containing pairs of honey-beans cell (hexagons for planar and tree-hexagons for spatial) that are neighbors in a 6/20-connected sense. The weight of each edge $e = (h_i, h_j)$ is denoted by $w(e)$, or similarly by $w(h_i, h_j)$, and it represents the dissimilarity between neighboring elements “ h_i ” and “ h_j ” in a some feature space. Components of an image represent compact regions containing pixels/voxels with similar properties. Thus, the set V of vertices of the graph G is partitioned into disjoint sets, each subset representing a distinct visual object of the initial image.

As in other graph-based approaches Burdescu et al. (2009) we use the notion of segmentation of the set V . A segmentation, S , of V is a partition of V , such that each component $C \in S$ corresponds to a connected component in a spanning sub-graph $G_S = (V, E_S)$ of G , with $E_S \subseteq E$.

The set of edges $E - E_S$ that are eliminated connect vertices from distinct components. The common boundary between two connected components $C', C'' \in S$ represents the set of edges connecting vertices from the two components:

$$cb(C', C'') = \{(h_i, h_j) \in E \mid h_i \in C', h_j \in C''\} \quad (3)$$

The set of edges $E - E_S$ represents the boundary between all components in S . This set is denoted by $bound(S)$ and it is defined as follows:

$$bound(S) = \bigcup_{C', C'' \in S} cb(C', C''). \quad (4)$$

In order to simplify notations throughout the paper we use C_i to denote the component of a segmentation S that contains the vertex $h_i \in V$.

We use the notions of segmentation “too fine” and “too coarse” as defined in Felzenszwalb & Huttenlocher (2004) that attempt to formalize the human perception of salient visual objects from an image. A segmentation S is too fine if there is some pair of components $C', C'' \in S$ for which there is no evidence for a boundary between them. S is too coarse when there exist a proper refinement of S that is not too fine. The key element in this definition is the evidence for a boundary between two components.

The goal of a segmentation method is to determine a proper segmentation, which represent visual objects from an image.

Definition 1 Let $G = (V, E)$ be the undirected planar/spatial graph constructed on the hexagonal/tree-hexagonal structure of an image, with $V = \{h_1, \dots, h_{|V|}\}$. A proper segmentation of V , is a partition S of V such that there exists a sequence $[S_i, S_{i+1}, \dots, S_{f-1}, S_f]$ of segmentations of V for which:

- $S = S_f$ is the final segmentation and S_i is the initial segmentation,
- S_j is a proper refinement of S_{j+1} (i.e., $S_j \subset S_{j+1}$) for each $j = i, \dots, f-1$,
- segmentation S_j is too fine, for each $j = i, \dots, f-1$,
- any segmentation S_1 such that $S_f \subset S_1$, is too coarse,
- segmentation S_f is neither too coarse nor too fine.

We present a unified framework for image segmentation and contour extraction that uses a virtual hexagonal structure defined on the set of the image pixels. This proposed graph-based segmentation method is divided into two different steps: (1) a pre-segmentation step that produces a maximum spanning tree of the connected components of the triangular grid graph constructed on the hexagonal structure of the input image; and (2) the final segmentation step that produces a minimum spanning tree of the connected components, representing the visual objects by using dynamic weights based on the geometric features of the regions (Stanescu et al. 2011).

Each hexagon from the hexagonal grid contains eight pixels: six pixels from the frontier and two interior pixels. Because square pixels from an image have integer values as coordinates we always select the left pixel from the two interior pixels to represent with approximation the gravity center of the hexagon, denoted by the pseudo-gravity center. We use a simple scheme of addressing for the hexagons of the hexagonal grid that encodes the spatial location of the pseudo-gravity centers of the hexagons as presented in Fig. 9.3 (for planar image).

Let $w \times h$ the dimension of the initial image. Given the coordinates $\langle h, c \rangle$ of a pixel “p” from the input image, we use the linear function, $ip_{w,h}(\langle l, c \rangle) = (l-1)w + c$, in order to determine an unique index for the pixel.

Let “ps” be the sub-sequence of the pixels from the sequence of the pixels of the initial image that correspond to the pseudo-gravity center of hexagons, and “hs” the sequence of hexagons constructed over the pixels of the initial image. For each pixel “p” from the sequence “ps” having the coordinates $\langle h, c \rangle$, the index of the corresponding hexagon from the sequence “hs” is given automatically by system. Equation for the hexagons is linear and it has a natural order induced by the sub-sequence of pixels representing the pseudo-gravity center of hexagons. Relations allow us to uniquely determine the coordinates of the pixel representing the pseudo-gravity center of a hexagon specified by its index (its address). Each hexagon represents an elementary item and the entire virtual hexagonal structure represents a triangular grid graph, $G = (V, E)$, where each hexagon “h” in this structure has a corresponding vertex $v \in V$. The set E of edges is constructed by connecting hexagons that are neighbors in a 6-connected sense. The vertices of this graph correspond to the pseudo-gravity centers of the hexagons from the hexagonal grid and the edges are straight lines connecting the pseudo-gravity centers of the neighboring hexagons, as presented in Fig. 9.3.

There are two main advantages when using hexagons instead of pixels as elementary pieces of information:

1. The amount of memory space associated with the graph vertices is reduced. Denoting by “np” the number of pixels of the initial image, the number of the resulted hexagons is always less than $np/4$, and thus the cardinal of both sets V and E is significantly reduced;
2. The algorithms for determining the visual objects and their contours are much faster and simpler in this case. Many of these algorithms are “borrowed” from graph sets Cormen et al. (1990).

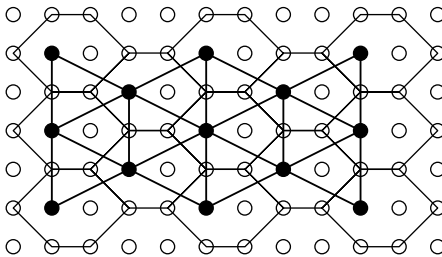


Fig. 9.3: The triangular grid graph constructed on the pseudo-gravity centers of the hexagonal grid.

We associate to each hexagon “h” from V two important attributes representing its dominant color and the coordinates of its pseudo-gravity center, denoted by “g(h).” The dominant color of a hexagon is denoted by “c(h)” and it represents the color of the pixel of the hexagon which has the minimum sum of color distance to the other seven pixels. Each hexagon “h” in the hexagonal grid is thus represented by a single point, g(h), having the color c(h). By using the values g(h) and c(h) for each hexagon information related to all pixels from the initial image is taken into consideration by the segmentation algorithm.

Our segmentation algorithm starts with the most refined segmentation, $S_0 = \{\{h_1\}, \dots, \{h_{|V|}\}\}$ and it constructs a sequence of segmentations until a proper segmentation is achieved. Each segmentation S_j is obtained from the segmentation S_{j-1} by merging two or more connected components for there is no evidence for a boundary between them. For each component of a segmentation a spanning tree is constructed; thus for each segmentation we use an associated spanning forest.

The evidence for a boundary between two components is determined taking into consideration some features in some model of the image. When starting, for a certain number of segmentations the only considered feature is the color of the regions associated to the components and in this case we use a color-based region model. When the components became complex and contain too much hexagons/tree-hexagons, the color model is not sufficient and hence geometric features together with color information are considered. In this case we use a syntactic based (with a color-based region) model for regions. In addition, syntactic features bring *supplementary* information for merging similar regions in order to determine salient objects. Despite of the majority of the segmentation methods our method do not require any parameter to be chosen or tuned in order to produce a better segmentation and thus our method is totally adaptive. The entire approach is fully *unsupervised* and does not need a priori information about the image scene (Burdescu et al. 2011; Brezovan et al. 2010).

For the sake of simplicity, we will denote this region model as a syntactic-based region model.

As a consequence, we split the sequence of all segmentations,

$$S_{if} = [S_0, S_1, \dots, S_{k-1}, S_k], \quad (5),$$

in two different subsequences, each subsequence having a different region model,

$$S_i = [S_0, S_1, \dots, S_{t-1}, S_t], \text{ and } S_f = [S_t, S_{t+1}, \dots, S_{k-1}, S_k], \quad (6),$$

where S_i represents the color-based segmentation sequence, and S_f represents the syntactic-based segmentation sequence.

The final segmentation S_t in the color-based model is also the initial segmentation in the syntactic-based region model.

For each sequence of segmentations we develop a different algorithm (Stanescu et al. 2011). Moreover, we use a different type of spanning tree in each case: a maximum spanning tree in the case of the color-based segmentation, and a minimum spanning tree in the case of the syntactic-based segmentation (Cormen, Leiserson & Rivest 1990). More precisely our method determines two sequences of forests of spanning trees,

$$F_i = [F_0, F_1, \dots, F_{t-1}, F_t], \text{ and } F_f = [F_{t'}, F_{t'+1}, \dots, F_{k'-1}, F_{k'}], \quad (7)$$

each sequence of forests being associated with a sequence of segmentations.

The first forest from F_i contains only the vertices of the initial graph, $F_0 = (V, \emptyset)$, and at each step some edges from E are added to the forest $F_1 = (V, E_1)$ to obtain the next forest, $F_{i+1} = (V, E_{i+1})$. The forests from F_i contain maximum spanning trees and they are determined by using a modified version of Kruskal’s algorithm (Cormen, Leiserson & Rivest 1990), where at each step the heaviest edge (u,v) that leaves the tree associated to “ u ” is added to the set of edges of the current forest.

The second subsequence of forests that correspond to the subsequence of segmentations S_f contains forests of minimum spanning trees and they are determined by using a modified form of Boruvka’s algorithm. This sequence uses as input a new graph, $G' = (V', E')$, which is extracted from the last forest, $F_{t'}$, of the sequence F_i . Each vertex “ v ” from the set V' corresponds to a component C_v from the segmentation S_t (i.e., to a region determined by the previous algorithm). At each step the set of new edges added to the current forest are determined by each tree T contained in the forest that locates the lightest edge leaving T . The first forest from F_f contains only the vertices of the graph G' , $F_{t'} = (V', \emptyset)$.

In this section we focus on the definition of a logical predicate that allow us to determine if two neighboring regions represented by two components, $C_{i'}$ and $C_{i''}$, from a segmentation S_t can be merged into a single component $C_{i'+1}$ of the segmentation S_{t+1} . Two components, $C_{i'}$ and $C_{i''}$, represent neighboring (adjacent) regions if they have a common boundary:

$$\begin{aligned} \text{adj}(C_{i'}, C_{i''}) &= \text{true} && \text{if } \text{cb}(C_{i'}, C_{i''}) \neq \emptyset, \\ \text{adj}(C_{i'}, C_{i''}) &= \text{false} && \text{if } \text{cb}(C_{i'}, C_{i''}) = \emptyset \end{aligned} \quad (8)$$

We use a different predicate for each region model, color based and syntactic-based, respectively.

$$\text{PED}(e, u) = [\text{wR}(\text{Re}-\text{Ru})^2 + \text{wG}(\text{Ge}-\text{Gu})^2 + \text{wB}(\text{Be}-\text{Bu})^2]^{1/2} \quad (9)$$

where the weights for the different color channels, w_R , w_G , and w_B verify the condition

$$w_R + w_G + w_B = 1.$$

Based on the theoretical and experimental results on spectral and real world data sets, Stanescu et al. (2011) is concluded that the PED distance with weight-coefficients ($w_R = 0.26$, $w_G = 0.70$, $w_B = 0.04$) correlates significantly higher than all other distance measures including the angular error and Euclidean distance.

In the color model regions are modeled by a vector in the RGB color space. This vector is the mean color value of the dominant color of hexagons/tree-hexagons belonging to the regions. There are many existing systems for arranging and describing colors, such as RGB, YUV, HSV, LUV, CIELAV, Munsell system, etc. (Billmeyer & Salzman 1981). We have decided to use the RGB color space because it is efficient and no conversion is required. Although it also suffers from the non-uniformity problem where the same distance between two color points within the color space may be perceptually quite different in different parts of the space, within a certain color threshold it is still definable in terms of color consistency.

The evidence for a boundary between two regions is based on the difference between the internal contrast of the regions and the external contrast between them (Felzenszwalb & Huttenlocher 2004; Stanescu 2011). Both notions of internal contrast and external contrast between two regions are based on the dissimilarity between two such colors.

Let h_i and h_j representing two vertices in the graph $G = (V, E)$, and let $wcol(h_i, h_j)$ representing the color dissimilarity between neighboring elements h_i and h_j , determined as follows:

$$\begin{aligned} wcol(h_i, h_j) &= PED(c(h_i), c(h_j)) && \text{if } (h_i, h_j) \in E, \\ \text{and } wcol(h_i, h_j) &= \infty && \text{otherwise,} \end{aligned} \quad (10)$$

where $PED(e, u)$ represents the perceptual Euclidean distance with weight-coefficients between colors “ e ” and “ u ,” as defined by Equation (9), and $c(h)$ represents the mean color vector associated with the hexagons or tree-hexagon “ h .” In the color-based segmentation, the weight of an edge (h_i, h_j) represents the color dissimilarity, $w(h_i, h_j) = wcol(h_i, h_j)$.

Let S_1 be a segmentation of the set V .

We define the *internal contrast* or *internal variation* of a component $C \in S_1$ to be the maximum weight of the edges connecting vertices from C :

$$IntVar(C) = \max_{(h_i, h_j) \in C} (w(h_i, h_j)). \quad (11)$$

The internal contrast of a component C containing only one hexagon is zero: $IntVar(C) = 0$, if $|C| = 1$.

The *external contrast* or *external variation* between two components, $C, C'' \in S$ is the maximum weight of the edges connecting the two components:

$$ExtVar(C', C'') = \max(hi, hj) \in cb(C', C'') (w(hi, hj)). \quad (12)$$

We had chosen the definition of the external contrast between two components to be the maximum weight edge connecting the two components, and not to be the minimum weight, as in Felzenszwalb & Huttenlocher W (2004) because: (1) it is closer to the human perception (perception of maximum color dissimilarity); and (2) the contrast is uniformly defined (as maximum color dissimilarity) in the two cases of internal and external contrast.

The maximum internal contrast between two components, $C, C'' \in S$ is defined as follows:

$$IntVar(C', C'') = \max(IntVar(C'), IntVar(C'')), \quad (13)$$

The comparison predicate between two neighboring components C' and C'' (i.e., $adj(C', C'') = true$) determines if there is an evidence for a boundary between C' and C'' and it is defined as follows:

$$\begin{aligned} diffcol(C', C'') = true, & \text{ if } ExtVar(C', C'') > IntVar(C', C'') + \tau(C', C''), \\ diffcol(C', C'') = false, & \text{ if } ExtVar(C', C'') \leq IntVar(C', C'') + \tau(C', C''), \end{aligned} \quad (14)$$

with the the adaptive threshold $\tau(C', C'')$ given by

$$\tau(C', C'') = \tau / \min(|C'|, |C''|), \quad (15)$$

where $|C|$ denotes the size of the component C (i.e., the number of the hexagons or tree-hexagons contained in C) and the threshold “ τ ” is a global adaptive value defined by using a statistical model.

The predicate $diffcol$ can be used to define the notion of segmentation too fine and too coarse in the color-based region model.

Definition 2 Let $G = (V, E)$ be the undirected spatial graph constructed on the hexagons or tree-hexagonal structure of planar or spatial image and S by color-based segmentation of V . The segmentation S is too fine in the color-based region model if there is a pair of components $C', C'' \in S$ for which

$$adj(C', C'') = true \wedge diffcol(C', C'') = false.$$

Definition 3 Let $G = (V,E)$ be the undirected planar or spatial graph constructed on the hexagons or tree-hexagonal structure of planar or spatial image and S a segmentation of V . The segmentation S is too coarse if there exists a proper refinement of S that is not too fine.

We use the perceptual Euclidean distance with weight-coefficients (PED) as the distance between two colors.

Let $G = (V,E)$ be the initial graph constructed on the tree-hexagonal structure of a spatial image. The proposed segmentation algorithm will produce a proper segmentation of V according to the Definition 1. The sequence of segmentations, S_{if} , as defined by Equation (5), and its associated sequence of forests of spanning trees, F_{if} , as defined by Equation (7), will be iteratively generated as follows:

- The color-based sequence of segmentations, S_i , as defined by Equation (6), and its associated sequence of forests, F_i , as defined by Equation (7), will be generated by using the color-based region model and a maximum spanning tree construction method based on a modified form of the Kruskal's algorithm (Cormen, Leiserson & Rivest 1990).
- The syntactic-based sequence of segmentations, S_p , as defined by Equation (6), and its associated sequence of forests, F_p , as defined by Equation (7), will be generated by using the syntactic-based model and a minimum spanning tree construction method based on a modified form of the Boruvka's algorithm.

The general form of the segmentation procedure is presented in Algorithm 1.

Algorithm 1 Segmentation algorithm for planar images

1. **** Procedure** SEGMENTATION ($l, c, P, H, Comp$)
2. **Input** l, c, P
3. **Output** $H, Comp$
4. $H \leftarrow *CREATEHEXAGONALSTRUCTURE (l, c, P)$
5. $G \leftarrow *CREATEINITIALGRAPH (l, c, P, H)$
6. $*CREATECOLORPARTITION (G, H, Bound)$
7. $G' \leftarrow *EXTRACTGRAPH (G, Bound, th_g^k)$
8. $*CREATESYNTACTICPARTITION (G, G', th_g^k)$
9. $Comp \leftarrow *EXTRACTFINALCOMPONENTS (G')$
10. **End procedure**

The input parameters represent the image resulted after the pre-processing operation: the array P of the planar image pixels structured in “ l ” lines and “ c ” columns. The output parameters of the segmentation procedure will be used by the contour extraction procedure: the hexagonal grid stored in the array of hexagons H , and the array $Comp$ representing the set of determined components associated to the objects in the input image.

The global parameter threshold “ th_g^k ” is determined by using Algorithm 1.

The color-based segmentation and the syntactic-based segmentation are determined by the procedures CREATECOLORPARTITION and CREATESYNTACTICPARTITION, respectively.

The color-based and syntactic-based segmentation algorithms use the hexagonal structure H created by the function CREATEHEXAGONALSTRUCTURE over the pixels of the initial image, and the initial triangular grid graph G created by the function CREATEINITIALGRAPH. Because the syntactic-based segmentation algorithm uses a graph contraction procedure, CREATESYNTACTICPARTITION uses a different graph, G , extracted by the procedure EXTRACTGRAPH after the color-based segmentation finishes.

Both algorithms for determining the color-based and syntactic based segmentation use and modify a global variable (denoted by CC) with two important roles:

1. to store relevant information concerning the growing forest of spanning trees during the segmentation (maximum spanning trees in the case of the color-based segmentation, and minimum spanning trees in the case of syntactic based segmentation),
2. to store relevant information associated to components in a segmentation in order to extract the final components because each tree in the forest represents, in fact, a component in each segmentation S in the segmentation sequence determined by the algorithm.

In addition, this variable is used to maintain a fast disjoint set-structure in order to reduce the running time of the color based segmentation algorithm. The variable CC is an array having the same dimension as the array of hexagons H , which contains as elements objects of the class *Tree* with the following associated fields: (isRoot, parent, compIndex, frontier, surface, color)

The field “isRoot” is a boolean value specifying if the corresponding hexagon index is the root of a tree representing a component, and the field parent represents the index of the hexagon which is the parent of the current hexagon. The rest of fields are used only if the field “isRoot” is true. The field “compIndex” is the index of the associated component.

The field “surface” is a list of indices of the hexagons belonging to the associated component, while the field “frontier” is a list of indices of the hexagons belonging to the frontier of the associated component. The field color is the mean color of the hexagon colors of the associated component.

The procedure EXTRACTFINALCOMPONENTS determines for each determined component C of $Comp$, the set “sa(C)” of hexagons belonging to the component,

the set $sp(C)$ of hexagons belonging to the frontier, and the dominant color $c(C)$ of the component.

A potential user of an algorithm's output needs to know what types of incorrect/invalid results to expect, as some types of results might be acceptable while others are not. This called for the use of metrics that are necessary for potential consumers to make intelligent decisions.

This presents the characteristics of the error metrics defined in Martin et al. (2001). The authors proposed two metrics that can be used to evaluate the consistency of a pair of segmentations, where segmentation is simply a division of the pixels of an image into discrete sets. Thus a segmentation error measure takes two segmentations S_1 and S_2 as input and produces a real valued output in the range $[0-1]$ where zero signifies no error.

The process defines a measure of error at each pixel that is tolerant of refinement as the basis of both measures. A given pixel " p_i " is defined in relation to the segments in S_1 and S_2 that contain that pixel. As the segments are sets of pixels and one segment is a proper subset of the other, then the pixel lies in an area of refinement and therefore the local error should be zero. If there is no subset relationship, then the two regions overlap in an inconsistent manner. In such case, the local error should be non-zero.

Let \setminus denote set difference, and $|x|$ the cardinality of set " x ." If $R(S; p_i)$ is the set of pixels corresponding to the region in segmentation S that contains pixel p_i , the local refinement error is defined as in Stanescu et al. (2011):

$$E(S1, S2, p_i) = \frac{|R(S1, p_i) \setminus R(S2, p_i)|}{|R(S1, p_i)|}$$

Note that this local error measure is not symmetric. It encodes a measure of refinement in one direction only: $E(S1; S2; p_i)$ is zero precisely when S_1 is a refinement of S_2 at pixel " p_i ," but not vice versa. Given this local refinement error in each direction at each pixel, there are two natural ways to combine the values into an error measure for the entire image. Global consistency error (GCE) forces all local refinements to be in the same direction. Let " n " be the number of pixels:

$$GCE(S1, S2) = \frac{1}{n} \min \left\{ \sum_i E(S1, S2, p_i), \sum_i E(S2, S1, p_i) \right\},$$

Local consistency error (LCE) allows refinement in different directions in different parts of the image.


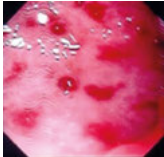
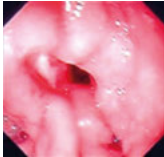
$$LCE(S1, S2) = \frac{1}{n} \sum_i \min \{ E(S1, S2, p_i), E(S2, S1, p_i) \}$$

As $LCE \leq GCE$ for any two segmentations, it is clear that GCE is a tougher measure than LCE. Martin et al. showed that, as expected, when pairs of human segmentations of the same image are compared, both the GCE and the LCE are low; conversely, when random pairs of human segmentations are compared, the resulting GCE and LCE are high.

9.6 Experimental results

Tables 9.1 and 9.2 show the images for which we and our colleagues provide experimental results (Mihai et al. 2011).

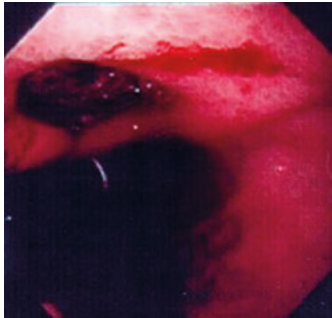
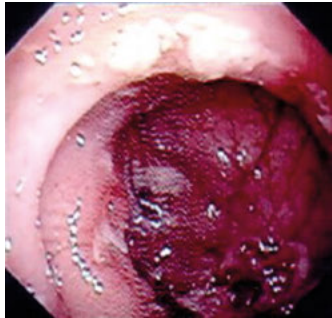
Tab. 9.1: Test Images and associated words.

Image	Diagnostic	Words
	Esophagitis	Inflammation, esophagus, esophageal diseases, gastrointestinal diseases, digestive system diseases
	Rectocolitis	Inflammation, rectum, colitis, gastroenteritis, gastrointestinal diseases
	Ulcer	Peptic ulcer, duodenal diseases, intestinal diseases, gastrointestinal diseases, digestive system diseases

For testing the annotation module, we have used a set of 2000 medical images: 1500 of images in the training set and 500 test images. In the table, below, we present the words assigned by the annotation system to some test images:

For testing the quality of our segmentation algorithm (by comparing GBOD with two other well-known algorithms – the local variation algorithm and the

Tab. 9.2: Images used in segmentation experiments.

Image number	
1	2
	

color-set back projection algorithm) the experiments were conducted using a database with 500 medical images of the digestive system, which were captured by an endoscope. The images were taken from patients having diagnoses such as *polyps, ulcers, esophagitis, colitis, and ulcerous tumors*.

For each image the following steps are performed by the application that we have created to calculate de GCE and LCE values:

1. Obtain the image regions using the color set back-projection segmentation – CS
2. Obtain the image regions using the local variation algorithm (LV)
3. Obtain the image regions using the graph-based object detection – GBOD
4. Obtain the manually segmented regions – MS
5. Store these regions in the database
6. Calculate GCE and LCE
7. Store these values in the database for later statistics

In Tab. 9.3 can be seen the number of regions resulted from the application of the segmentation.

Tab. 9.3: The number of regions detected for each algorithm.

Img. no	CS	LV	GBOD	MS
1	9	5	3	4
2	8	7	2	3

In Tab. 9.4 are presented the GCE values calculated for each algorithm.

Tab. 9.4: GCE values calculated for each algorithm.

Img. no	GCE-CS	GCE-GBOD	GCE-LV
1	0.18	0.09	0.24
2	0.36	0.10	0.28

In Tab. 9.5 are presented the LCE values calculated for each algorithm.

Tab. 9.5: LCE values calculated for each algorithm.

Image no	LCE-CS	LCE-GBOD	LCE-LV
1	0.11	0.07	0.15
2	0.18	0.12	0.17

Figures 9.4 and 9.5 present the regions resulted from manual segmentation and from the application of the segmentation algorithm presented above for images displayed in Tab. 9.2.

If a different segmentation algorithm arises from different perceptual organizations of the scene, then it is fair to declare the segmentations inconsistent. If, however, the segmentation algorithm is simply a refinement of the other, then the error should be small, or even zero. The error measures presented in the above tables are calculated in relation with the manual segmentation which is considered true segmentation. From Tabs. 9.3 and 9.4 it can be observed that the values for GCE and LCE are lower in the case of GBOD method. The error measures, for almost all tested images, have smaller values in the case of the original segmentation method, which employs a hexagonal structure defined based on the set of pixels.

Figure 9.6 presents the repartition of the 500 images from the database repartition on GCE values. The focal point here is the number of images on which the GCE value is under 0.5. In conclusion for GBOD algorithm, a number of 391 images (78%) obtained GCE values under 0.5. Similarly, for CS algorithm only 286 images (57%) obtained GCE values under 0.5. The segmentation based on LV method is close to our original algorithm: 382 images (76%) had GCE values under 0.5.



Fig. 9.4: The resulted regions for image number 1.

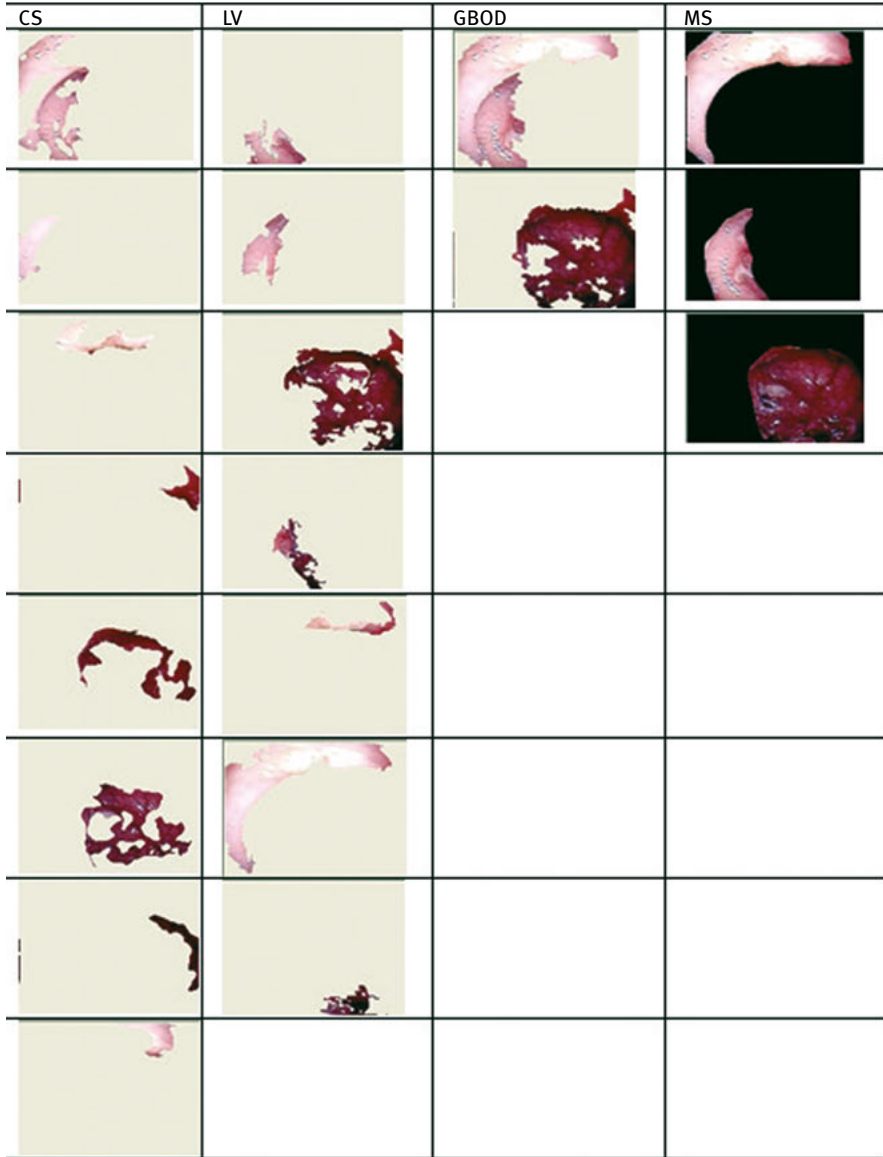


Fig. 9.5: The resulted regions for image number 2.

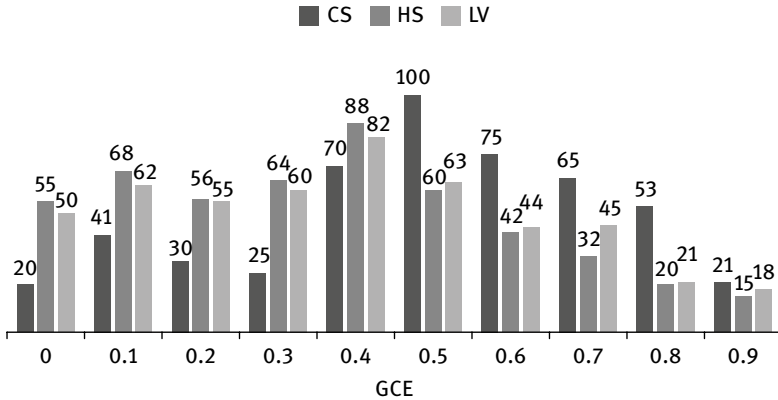


Fig. 9.6: Number of images relative to GCE values.

9.7 Conclusions

The testing scenario used by a medical doctor included the following steps: (1) A new image was obtained using the endoscope (planar and RGB image). (2) The image obtained was then processed by the annotation system and a set of words were suggested. (3) The doctor proceeded to analyze the words that were suggested along with the processed image. (4) The doctor concluded the assigned words were relevant to the image, and the system was as a *starting point* for the diagnostic process. Nevertheless, in order to establish and to validate a correct diagnosis of the patient's condition, additional medical investigation is needed, *inter alia*, medical tests and procedures as well as a comprehensive medical history. However, by having a large enough annotated dataset of images the system can correctly suggest the diagnosis, which was the main purpose of implementing our annotation system.

In this chapter, we evaluated three algorithms used to detect regions in endoscopic images: a clustering method (the color set back-projection algorithm), as well as two other methods of segmentation based on graphs: (1) the local variation algorithm; and (2) our original segmentation algorithm (GBOD). Our method is based on a hexagonal structure defined on the set of image pixels. The advantage of using a virtual hexagonal network superimposed over the initial image pixels is that it reduces the execution time and the memory space used, without losing the initial resolution of the image.

Furthermore, because the error measures for segmentation using GBOD method are lower than for color set back-projection and local variation segmentation, we can infer that the proposed segmentation method based on a hexagonal structure is more efficient. Our experimental results show that the original GBOD segmentation method is a good refinement of the manual segmentation.

In comparison to other segmentation methods, our algorithm is able to adapt and does not require either parameters for establishing the optimal values or sets of training images to set parameters. More specifically, following our application of the *three* algorithms used to detect regions in radiographic images, we saw direct evidence of the adaptation of our methods when we *compared* the correctness of the image segments to the assigned words. Our study findings conclusively showed that concerning the endoscopic database, *all* the algorithms have the ability to produce segmentations that comply with the manual segmentation made by a medical expert. As part of our experiment, we used a set of segmentation error measures to evaluate the accuracy of our annotation model.

Medical images can be described properly only by using a set of specific words. In practice, this constraint can be satisfied by the usage of ontology. Several design criteria and development tools were presented to illustrate the means available for creating and maintaining ontology. All in all, building ontology for representing medical terminology systems is a difficult task that requires a profound analysis of the structure and the concepts of medical terms, but necessary in order to solve diagnostic problems that frequently occur in the day-to-day practice of medicine. Medical Subject Headings (MeSH) (Martin et al. 2001) is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it can also serve as a thesaurus that can be used to assist in a variety of searching tasks. Created and updated by the United States National Library of Medicine (NLM), it is used by the MEDLINE/PubMed article database and by NLM's catalog of book holdings. In MEDLINE/PubMed, every journal article is indexed with some 10–15 headings or subheadings, with one or two of them designated as major and marked with an asterisk. When performing a MEDLINE search via PubMed, entry terms are automatically translated into the corresponding descriptors. The NLM staff members who oversee the MeSH database continually revise and update its vocabulary. *In essence, image classification and automatic image annotation might be treated as one of the effective solutions that enable keyword-based semantic image retrieval.* Undoubtedly, the importance of automatic image annotation has increased with the growth of digital images collections, as it allows indexing, retrieving, and understanding of large collections of image data. We have presented the results of our system created for evaluating the performance of annotation and retrieval (semantic-based and

content-based) tasks (Stanescu et al. 2011). Our present system provides support for *all* steps that are required for evaluating the tasks mentioned above, including data import, knowledge storage and representation, knowledge presentation, and means for task-evaluation.

References

- Adamek, T., O'Connor, N. E. & Murphy, N. (2005) 'Region-based segmentation of images using syntactic visual features'. In *IMVIP 2005 – 9th Irish Machine Vision and Image Processing Conference*, Northern Ireland.
- Allène, C., Audibert, J.-Y., Couprie, M. & Keriven, R. (2010) 'Some links between extremum spanning forests, watersheds and min-cuts', *Image Vision Comput*, 28(10): 1460–1471.
- Barnard, K., Duygulu, P., De Freitas, N., Forsyth, D., Blei, D. & Jordan, M. I. (2003) 'Matching words and pictures', *J Mach Learn Res*, 3:1107–1135.
- Baoli, L., Ernest, V. G. & Ashwin, R. (2007) *Semantic Annotation and Inference for Medical Knowledge Discovery*, NSF Symposium on Next Generation of Data Mining (NGDM-07), Baltimore, MD.
- Billmeyer, F. & Salzman, M. (1981) 'Principles of Color Technology'. New York: Wiley.
- Blei, D. & Jordan, M. I. (2003) Modeling annotated data. In *Proceedings of the 26th Intl. ACM SIGIR Conf.*, pp. 127–134.
- Brezovan, M., Burdescu, D., Ganea, E. & Stanescu, L. (2010) An Adaptive Method for Efficient Detection of Salient Visual Object from Color Images. *Proceedings of the 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2346–2349.
- Brown, P., Pietra, S. D., Pietra, V. D. & Mercer, R. (1993) 'The mathematics of statistical machine translation: Parameter estimation', *In Computational Linguistics*, 19(2):263–311.
- Burdescu, D. D., Brezovan, M., Ganea, E. & Stanescu, L. (2009) *A New Method for Segmentation of Images Represented in a HSV Color Space*. Springer: Berlin/Heidelberg.
- Burdescu, D. D., Brezovan, M., Ganea, E. & Stanescu, L. (2011) 'New algorithm for segmentation of images represented as hypergraph hexagonal-grid', *IbPRIA*, 2011:395–402.
- Burdescu, D. D., Mihai, G. Cr., Stanescu, L. & Brezovan, M. (2013) 'Automatic image annotation and semantic based image retrieval for medical domain', *Neurocomputing*, 109:33–48, ISSN: 0925-2312.
- Catherine, E. C., Xenophon, Z. & Stelios, C. O. (1997) 'I2Cnet Medical image annotation service', *Med Inform, Special Issue*, 22(4):337–347.
- Cormen, T., Leiserson, C. & Rivest, R. (1990) *Introduction to Algorithms*. Cambridge, MA: MIT Press.
- Daniel, E. (2003) *OXALIS: A Distributed, Extensible Ophthalmic Image Annotation System*, Master of Science Thesis.
- Duygulu, P., Barnard, K., de Freitas, N. & Forsyth, D. (2002) 'Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary', *In Seventh European Conf. on Computer Vision*, pp. 97–112.
- Felzenszwalb, P. & Huttenlocher, W. (2004) 'Efficient graph-based image segmentation', *Int J Comput Vis*, 59(2):167–181.

- Grundmann, M., Kwatra, V., Han, M. & Essa., I. (2010) Efficient hierarchical graph-based video segmentation. In *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR 2010)*.
<http://www.ncbi.nlm.nih.gov/pubmed>
<http://www.nlm.nih.gov/>
<http://www.nlm.nih.gov/mesh/filelist.html>
<http://www.nlm.nih.gov/mesh/meshrels.html>
http://www.nlm.nih.gov/mesh/2010/mesh_browser/MeSHtree.html
<http://www.obofoundry.org/>
<http://www.topicmaps.org/>
http://en.wikipedia.org/wiki/Medical_Subject_Headings
- Igor, F. A., Filipe, C., Joaquim, F., Pinto, da C. & Jaime, S. C. (2010) Hierarchical Medical Image Annotation Using SVM-based Approaches. In *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine*.
- Jeon, J., Lavrenko, V. & Manmatha, R. (2003) Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In *Proceedings of the 26th International ACM SIGIR Conference*, pp. 119–126.
- Jin, R., Chai, J. Y. & Si, L. (2004) 'Effective automatic image annotation via a coherent language model and active learning', In *ACM Multimedia Conference*, pp. 892–899.
- Kohli, P., Silberman, N., Hoiem, D. & Fergus, R. (2012) Indoor segmentation and support inference from RGBD images, in ECCV.
- Lavrenko, V., Manmatha, R. & Jeon, J. (2004) A Model for Learning the Semantics of Pictures. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems, NIPS'03*.
- Li, J. & Wang, J. (2003) 'Automatic linguistic indexing of pictures by a statistical modeling approach'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.
- Liew, A. W.-C & Yan, H. (2005) 'Computer Techniques for Automatic Segmentation of 3D MR Brain Images'. In: *Medical Imaging Systems Technology: Methods in Cardiovascular and Brain Systems* (Vol. 5) Cornelius, T. Leondes (ed.) pp. 307–359. Singapore, London: World Scientific Publishing.
- Martin, D., Fowlkes, C., Tal, D. & Malik, J. (2001) A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics, IEEE (ed.), *Proceedings of the Eighth International Conference on Computer Vision (ICCV-01)*, Vancouver, British Columbia, Canada, vol. 2, 416–425.
- Mihai, G. Cr., Stanescu, L., Burdescu, D. D., Stoica-Spahiu, C., Brezovan, M. & Ganea, E. (2011) Annotation System for Medical Domain – Advances in Intelligent and Soft Computing, vol. 87, pg. 579–587, ISSN 1867-6662, Berlin, Heidelberg: Springer-Verlag.
- Mori, Y., Takahashi, H. & Oka, R. (1999) Image-to-word transformation based on dividing and vector quantizing images with words. In *MISRM'99 First Intl. Workshop on Multimedia Intelligent Storage and Retrieval Management*.
- Ojala, T., Pietikainen, M. & Harwood, D. (1996) 'A comparative study of texture measures with classification based on feature distributions', *Pattern Recogn*, 29(1):51–59.
- Peng, H., Long, F. & Myers, E. W. (2009) 'VANO: a volume-object image annotation system', *Bioinformatics*, 25(5):695–697.
- Stanescu, L., Burdescu, D. D., Brezovan, M. & Mihai, C. R. G. (2011) *Creating New Medical Ontologies for Image Annotation*, New York: Springer-Verlag.

- Stumme, G. & Madche, A. (2001) FCA-Merge: Bottom-up merging of ontologies. In *(IJCAI'01) Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Vol. 1, pp. 225–230.
- Tousch, A. M., Herbin, S. & Audibert, J. Y. (2012) ‘Semantic hierarchies for image annotation: A survey’, *Pattern Recogn* 45(1):333–345.

Randi Karlsen, Jose Enrique Borrás Morell, Johan Gustav Bellika and Vicente Traver Salcedo

10 Helping patients in performing online video search: evaluating the importance of medical terminology extracted from MeSH and ICD-10 in health video title and description

Abstract: Huge amounts of health-related videos are available on the Internet, and health consumers are increasingly searching for answers to their health problems and health concerns by availing themselves of web-based video sources. However, a critical factor in identifying relevant videos based on a textual query is the accuracy of the *metadata* with respect to video content. This chapter focuses on how reputable health videos providers, such as hospitals and health organizations, describe diabetes-related video content and the frequency with which they use standard terminology found in medical thesauri. In this study, we compared video title and description to medical terms extracted from the MeSH and ICD-10 vocabularies, respectively. We found that only a small number of videos were described using medical terms (4% of the videos included an exact ICD-10 term; and 7% an exact MeSH term). Furthermore, of all those videos that used medical terms in their title/description, we found an astonishingly low *variety* of diabetes-related medical terms used. For example, the video titles and descriptions brought up only 2.4% of the ICD-10 terms and 4.3% of MeSH terms, respectively. These figures give one pause to think as to how many useful health videos are haplessly eluding online patient search because of the sparse use of appropriate terms in titles and descriptions. Thus, no one would deny that including medical terms in video title and description is useful to patients who are searching for relevant health information. Adopting good practices for titling and describing health-related videos may similarly help producers of YouTube videos to identify and address the gaps in the delivery of informational resources that patients need to be able to monitor their own health. Sadly, as the situation is now, neither patients nor producers of health videos are able to explore the collection of online materials in the same systematic manner as the medical professional explores medical domains using MEDLINE. Why can we not have the same level of rigorous and systematic curating of patient-related health videos as we have for other medical content on the web?

10.1 Introduction

A huge amount of health information is available on the Internet, which has become a major source of information concerning many aspects of health (AlGhamdi & Moussa 2012; Griffiths et al. 2012). People are using the Internet to search for information about specific diseases or symptoms, read someone else's commentary or experience about health or medical issues, watch online health videos, consult online reviews of drug or medical treatments, search for others who might have health concerns similar to theirs and follow personal health experiences through blogs (de Boer, Versteegen & van Wijhe 2007; Powell et al. 2011; Fox 2011b).

Health information on the Internet comes from many different sources, including hospitals, health organizations, government, educational institutions, for-profit actors and private persons. However, the general problem of information overload makes it difficult to find relevant, good-quality health information on the Internet (Purcell, Wilson & Delamothe 2002; Mishoe 2008). Adding to this problem is that many websites have inaccurate, missing, obsolete, incorrect, biased or misleading information, often making it difficult to discern between veritable information and specious information found on the Web (Steinberg et al. 2010; Briones et al. 2012; Singh, Singh & Singh 2012; Syed-Abdul et al. 2013).

An important factor for information trustworthiness is the credibility of the information source (Freeman & Spyridakis 2009). Users are for example much more likely to trust health information published or authored by physicians or major health institutions (Dutta-Bergman 2003; Moturu, Lui & Johnson 2008; Bermudez-Tamayo et al. 2013) than information circulating in the blogosphere by other patients. Thus, users show greater interest in health information emanating from hospitals and health organizations (such as The American Diabetes Foundation and Diabetes UK) because these sources are considered more credible than the average health information put out on the web by other patients.

In this chapter, we focus primarily on health information provided through videos, and look at ways to provide health consumers with relevant videos to satisfy their informational needs. One of the most important factors in identifying relevant videos based on a textual query is the organization of metadata for cataloguing video material. For example, making sure that both the video title and the description of the video itself are accurate with respect to the content of the video. Given the importance of metadata for proper classification and retrieval of key health-related videos we take a close look at how reputable health videos providers, such as hospitals and health organizations, describe their video content through the use of medical terminology. In this study, we compare video title and description to medical terms extracted from the MeSH and ICD-10 vocabularies.

For practicality we chose to base our study on a narrowly defined clinical topic. This study is based on health videos obtained from YouTube through textual search queries on diabetes-related issues. YouTube is today the most important video-sharing website on the Internet (Cheng, Dale & Liu 2008), and is increasingly being used to share health information offered by hospitals, organizations, government, companies and private users (Bennett 2011) YouTube social media tools allow users to easily upload, view and share videos, and enable interaction by letting users rate videos and post comments. A persistent problem, however, is the difficulty in finding relevant health videos from credible sources such as hospitals and health organizations, given that search engine optimization tends to favor YouTube videos, thereby giving higher ranking favors to content stemming from popular sources (channels) than from more official sites. This should not be surprising given that unlike patient-generated YouTube videos, hospital and other healthcare organization videos do not readily benefit from social media interaction through likes/dislikes and comments, and for this, and other reasons, they tend to appear lower in the ranked list of online search results than the video material that is produced by lay sources.

We conducted the study by first issuing a number of diabetes-related queries to YouTube and identifying videos coming from more official healthcare organization sources. Title and description of those videos were checked for medical terms as found in the ICD-10 and MeSH vocabularies, and the prevalence of medical terms was determined.

The study was intended to answer the following questions: To what extent does title and description of diabetes health videos contain medical terms, as provided through the MeSH and ICD-10 vocabularies? Which medical terms are used in video title and description?

The chapter is broken down into several sections. Following the introductory section, above, Section 2 describes the MeSH and ICD-10 vocabularies, how test videos were obtained and how the study was conducted. Section 3 presents the results of the study, while Section 4 discusses the findings and how medical terminology in video title and description can be useful. Section 5 provides a conclusion to the chapter.

10.2 Data and methods

10.2.1 Obtaining video data

The videos used in this test were collected over a period of 44 days. We chose these parameters for the duration of our study because this 6 week or month and

a half period constitutes a time frame that is a long enough period for extracting useful conclusions. Thus we began our study on the 28th of February 2013 and ended on April 12th of that year. For each day, 19 queries focusing on different aspects of diabetes were issued to the YouTube web site. Each day, the top-500 ranked videos from each query were examined in order to identify videos coming from credible sources.

Through our study of health videos on YouTube we identified credible YouTube channels, such as hospitals and health organizations, and organized them into white-lists¹ of channels. The Health Care Social Media List started by Ed Bennett² was used as an initial white-list for credible channels, which we expanded with more channels that were identified during our studies (Karlsen et al. 2013).

Since users also seek information from their peers, we also identified videos coming from users that we classified as active in publishing diabetes-related videos. The generous availability of peer-to-peer healthcare video material demonstrates the fact that not only do patients and their caregivers have knowledge and experiences that they want to share but that patients themselves eagerly seek such information from their peers (Fox 2011a). To recognize this information need and to, likewise, investigate the use of medical terms in user-provided videos, we identified, in a third white-list, channels of active users that predominantly produced diabetes videos.

Our white-lists contained a total of 699 channels, where 651 were hospitals, 30 were organizations, and 18 were active users. The 19 queries used in the study all included the term “diabetes” and focused on different aspects concerning the disease. We used queries such as “diabetes a1c,” “diabetes glucose,” “diabetes hyperglycemia” and “diabetes lada.”

To execute the first stage of the project, which was the information-collection stage, we implemented a system that for each day of the study automatically issued the 19 queries (with an English anonymous profile). For each query, the system extracted information about the top 500 YouTube results, and identified new videos from white-listed channels that were included in our set of test videos. After the 44-days of the study, we had a total of 1380 unique diabetes-related videos from hospitals, health organizations, and active users. The title and description of each video were extracted and compared against medical terms in order to detect videos where its metadata contained medical terminology.

¹ The hospital white-list contains a list of YouTube channel identifiers, each identifying a separate hospital. The health organizations' white-list contains channel identifiers, each identifying a distinct health organization.

² <http://network.socialmedia.mayoclinic.org/hcsml-grid/>

10.2.2 Detecting medical terms in video title and/or description

The experiments were carried out using two different sets of medical terms: one from the ICD-10 database,³ and a second set from the MeSH database.⁴ The medical vocabulary used was the result of searching the online databases of ICD-10 and MeSH with the phrase “*diabetes mellitus.*” For ICD-10 we retrieved 167 medical terms related to diabetes, while MeSH returned 282 diabetes related terms. A term is either a single word or a phrase consisting of two or more words.

The set of terms were selected from the MeSH vocabulary (in the following denoted as MeSH terms set) and from the ICD-10 vocabulary (in the following denoted as ICD-10 terms set). We further distinguish between exact terms and partial terms. An *exact term* is a complete medical term as given in the ICD-10 or MeSH vocabulary. A *partial term* is a single word and a subset of an exact term (which may constitute a phrase). For example, “*neuropathy*” is a partial term of the exact term “*diabetic autonomic neuropathy.*”

To generate lists of partial terms from ICD-10 and MeSH vocabulary, we first removed stop words, such as “of” “the” “in” “with” “and” (Baeza-Yates and Ribeiro-Neto 2011). In addition, since diabetes is the general topic of the selected videos and vocabulary terms, we choose to discard the most common terms (“diabetes” “diabetic”) from our list of partial terms. This was done to focus the attention on more specific medical terms related to diabetes. We also did a manual revision of the partial terms, to discard terms that are not specifically related to the diabetes disease. This included terms such as (*diet, coma, complications, obese, latent, chemical, onset*), which were used in phrases such as “*diabetic coma*” and “*diabetic diet.*” After processing the medical phrases, we had two lists of partial medical terms, one for the ICD-10 terms set and the other for the MeSH terms set.

We used two analytic methods to identify medical terms in video title and description, where we focused on *exact term match* and *partial term match* to ICD-10 and MeSH terms set, respectively. Criteria for determining medical term prevalence in video title and description were as follows:

1. Exact match: The system detects a positive match when the video title/description contains an exact term from a vocabulary terms set.
2. Partial match: The system detects a positive match when the video title/description contains a partial term from a vocabulary terms set.

³ <http://apps.who.int/classifications/icd10/browse/2010/en>.

⁴ <http://www.ncbi.nlm.nih.gov/mesh>.

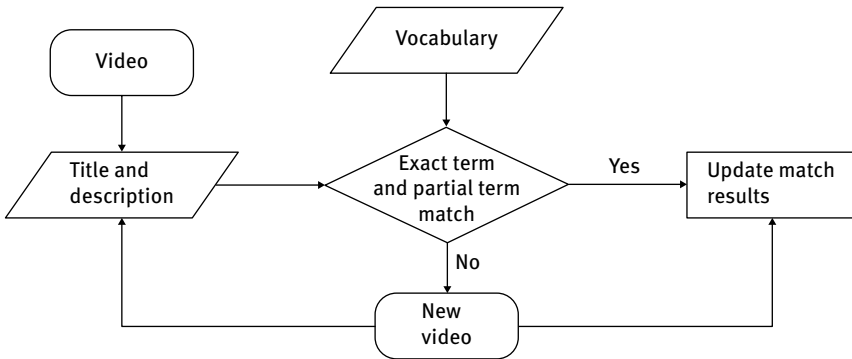


Fig. 10.1: Architecture for medical term detection.

Figure 10.1 shows how medical terms are detected in video title and description. Title and description of a video is, together with a vocabulary term set, given as input to a process that detects exact term matches and partial term matches for the video. If medical terms are detected, information about the video and the detected terms are stored. The system then continues with the next video, and searches for medical terms in the subsequent video.

We handle medical terms from ICD-10 and MeSH separately, meaning that the system only includes exact and partial terms from one of the vocabularies at any given time. Medical terms detection is thus executed twice for the videos: once for detecting ICD-10 terms; and subsequently, in the next execution of the system, for detecting MeSH terms. This was done in order to identify the impact different vocabularies had on medical terms detection.

10.2.3 Medical vocabularies

The International Classification of Diseases is maintained by WHO (World Health Organization) and is now in its tenth revision (ICD-10) since its first version in 1893. The purpose of ICD-10 is to provide a common foundation for definition of diseases and health conditions that allows the world to compare and share health information using a common language (WHO 2013). The classification is used to report and identify global health trends. The classification is used primarily by health workers. We have used the 10th Revision (ICD-10) to define exact and partial match with metadata description of the videos.

Medical Subject Headings (MeSH) is the National Library of Medicine’s controlled vocabulary thesaurus used for indexing articles for PubMed. MeSH is continuously updated with terms appearing in medical publications (NLM 2013). Terms like “Diabetes Mellitus” and even “Twitter Messaging” are defined in a hierarchical structure that permits searching at various specificity levels.

Together these vocabularies can be used to describe the relevant health condition covered by creators of some pieces of information while placing the piece(s) into a hierarchy of information that allows identification of the piece by healthcare information seekers.

10.3 Results

The test video collection contained 1380 distinct videos, including 270 hospital videos, 854 health organizations’ videos and 256 active users’ videos. The videos were uploaded from 73 hospital channels, 30 organization channels and 18 user channels. The videos’ title and description were compared against 167 ICD-10 terms and 282 MeSH terms.

Table 10.1 shows the results, in number of videos, from the two analytic methods: exact term match and partial term match. For both vocabularies a relatively high number of videos had a title and/or description including one or more terms that partially matched a term in the vocabulary. A low number of videos included exactly matching terms (3.9% for ICD-10 and 6.7% for MeSH), while around 40% of the videos did not include any medical terms at all.

Tab. 10.1: Final results. The number of videos with an exact match, partial match and no match to the ICD-10 and MeSH term sets (amount and percentage between parentheses).

	Exact match	Partial match	No match
ICD-10	54 (3.9%)	816 (59.1%)	510 (36.9%)
MeSH	92 (6.7%)	726 (52.6%)	562 (40.7%)

10.3.1 ICD-10 results

When comparing ICD-10 terms to video title and description, we found that 54 of the 1380 videos (4%) had a title/description with an exact term match to at least

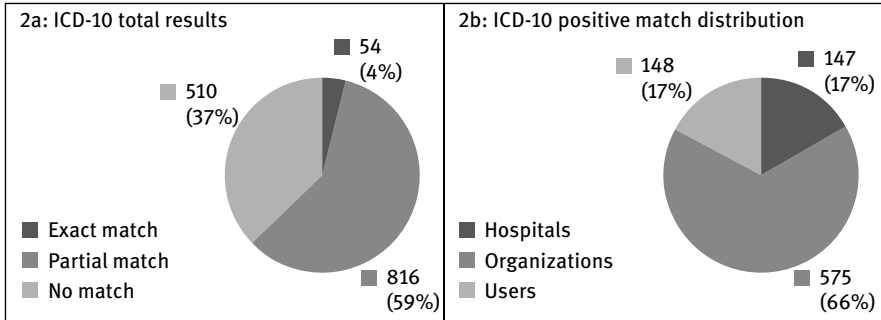


Fig. 10.2: Results of ICD-10 term matching, with total result over all videos (2a) and positive match distribution over hospital, health organization and active users videos (2b).

one ICD-10 term. For partial ICD-10 terms, we detected a positive match for 816 videos (59%), while 510 videos (37%) did not have any ICD-10 terms in its title and/or description altogether (see Fig. 10.2a).

When analyzing the distribution of positive result videos from hospitals, health organizations and active users, we found that the majority (66%) of videos with medical terms came from health organizations (see Fig. 10.2b), while videos from hospitals and active users had an equally low prevalence of ICD-10 terms (17%).

Figure 10.3 displays the total number of videos within each group (i.e., hospitals, health organizations, and active users) that had a title/description with an exact match (3a) or partial match (3b) to ICD-10 terms. As shown in Fig. 10.3, the videos produced by organizations contained the highest number of both

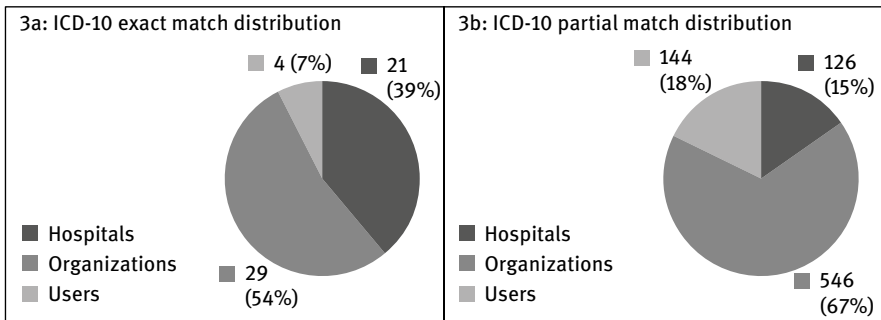


Fig. 10.3: Exact and partial match distribution over hospital, health organization and active users' videos.

exact match terms and partial match terms (54% and 67%, respectively) when compared to hospital and user-generated videos. For hospitals however, which came in second place, we noticed a significantly higher occurrence of exact match terms (39%), (more than twice as many of the partial match terms (15%) found in such videos), a mirror opposite of organizations whose partial match terms exceeded their exact match terms.

The relative proportion of videos from each group, with respect to exact and partial match, is displayed in Fig. 10.4. There we see that hospital videos had the largest relative proportion of exact match terms. Among the 270 hospital videos, 21 (7.8%) had an exact match with an ICD-10 term. The corresponding numbers for organizations and active users were 3.4% (29 of 854 videos) and 1.6%, (4 of 256 videos), respectively. The proportion of videos with a partial match to ICD-10 terms is much higher, with videos from health organizations coming out on top with 63.9%, followed successively by active users videos (56.3%) and hospital videos (46.7%).

During the analysis we detected a total number of 27 distinct ICD-10 terms. Figure 10.5 presents the most frequently used ICD-10 terms, including all terms that were used in more than five videos. Terms with an occurrence of five or less are represented as “other terms,” and include 15 terms with a total of 39 occurrences. Exact terms in Fig. 10.5 are “diabetic retinopathy,” “diabetic ketoacidosis” and “insulin-dependent.” The rest of the terms are partial ICD-10 terms. For a complete list of both exact and partial terms, see Tabs. 10.2 and 10.3, respectively.

In Fig. 10.5, we include the total number of ICD-10 term occurrences. This means that if a title/description includes the terms “type 1” and “glucose,” both terms are counted and represented in the numbers given in Fig. 10.5. We observe that four terms, “type 1,” “type 2,” “insulin” and “glucose,” have a much higher number of occurrences than the other terms. These four terms represent 81% of the total amount of ICD-10 term occurrences.

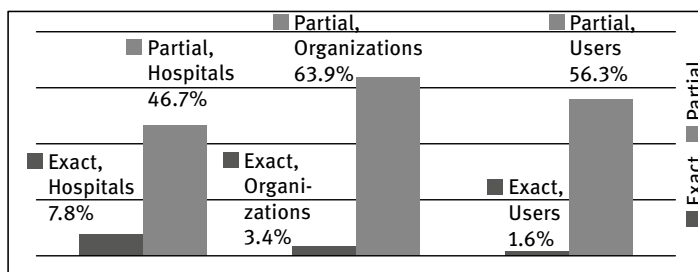


Fig. 10.4: Relative proportion of videos with medical terms.

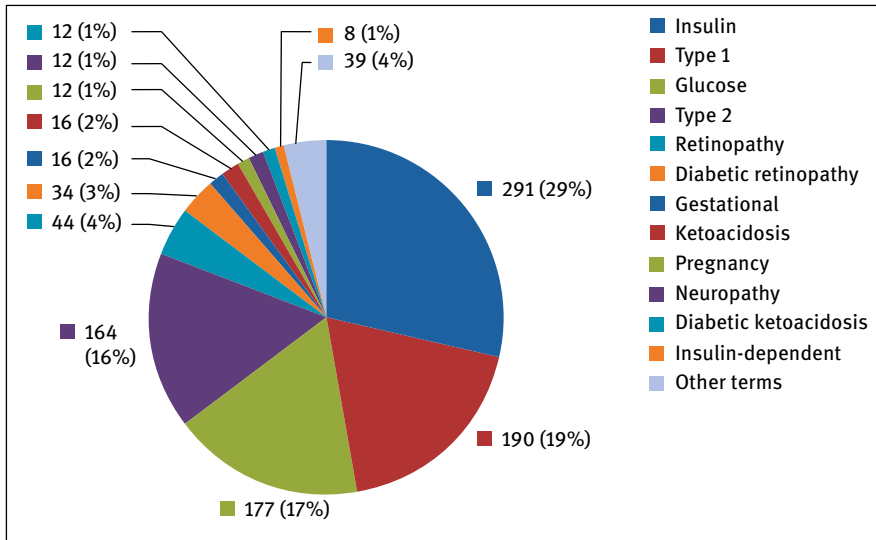


Fig. 10.5: The most frequently used ICD-10 terms in video title/descriptions.

10.3.2 MeSH Results

When comparing video title/description to MeSH terms, we found that 92 videos (7%) had an exact term match while 726 videos (52%) had a partial term match. Finally, there were 562 videos (41%) that did not have any MeSH terms in its title/description (see Fig. 10.6a).

We used the same analysis method as for ICD-10 terms, and present in Fig. 10.6b the positive match distribution over hospitals', organizations' and active users' videos, while in Fig. 10.7 we distinguish between exact match distribution and partial match distribution. For both ICD-10 and MeSH terms, we observe that hospital videos have a higher proportion of exact term matches compared with the proportion of partial term matches found in hospital videos.

The relative proportion of videos with MeSH terms (see Fig. 10.8), follows the same pattern as ICD-10 term matches (see Fig. 10.4), where hospitals have the highest relative proportion of exact term matches and organization have the highest relative proportion of partial term matches. A difference we noticed when looking for terms found in the MeSH and ICD-10 vocabularies is that the proportion of *exact* terms are for all groups higher when using the MeSH thesaurus as opposed to the ICD-10 thesaurus. In contrast, the proportion of *partial* terms for all groups were lower in the MeSH vocabulary when compared to ICD-10 vocabulary. This is why it is so important when doing this kind of research to distinguish

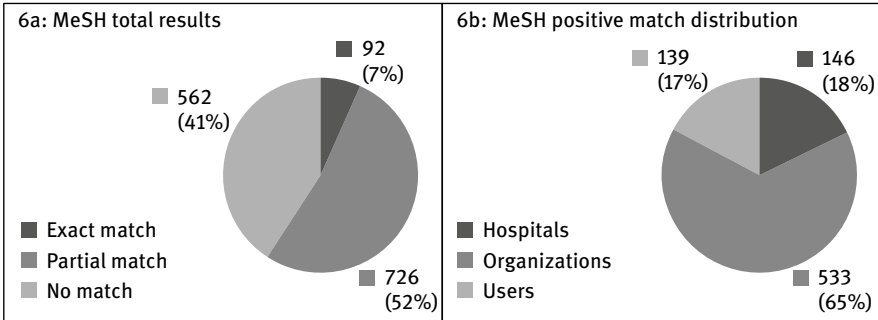


Fig. 10.6: Results of MeSH term matching, with total result over all videos and positive match distribution.

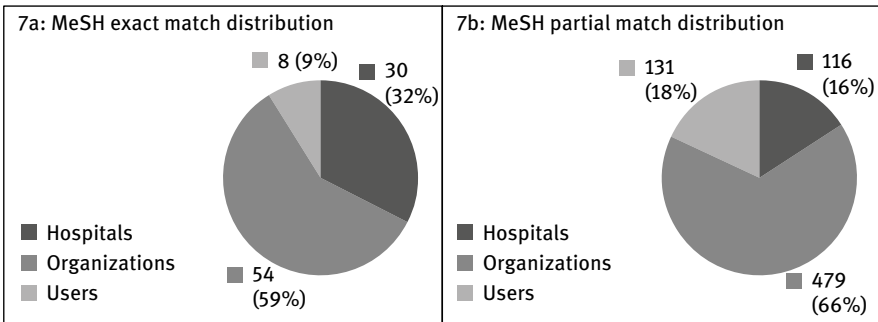


Fig. 10.7: MeSH exact and partial match distribution over hospital, health organization and active users videos.

between these two medical vocabularies as well as the kind of term that is found, indicating whether it is an exact or partial term.

We detected a total number of 45 distinct MeSH terms. Figure 10.9 presents the most frequently used MeSH terms, including all terms that are used in more than five videos. Terms with an occurrence of 5 or less are represented as “other terms,” and include 25 terms with a total of 66 occurrences. Exact match terms in Fig. 10.9 are “diabetic retinopathy,” “gestational diabetes,” “diabetic ketoacidosis” and “prediabetes.” The rest of the terms consist of partial MeSH terms. The complete list of both exact and partial MeSH terms used in the test videos are seen in Tabs. 10.2 and 10.3, respectively.

As was the case for ICD-10 terms, the four terms (*type 1*, *type 2*, *insulin*, and *glucose*) had a very high number of occurrences while the majority of terms had relatively few occurrences. Using MeSH, we see that four most common terms

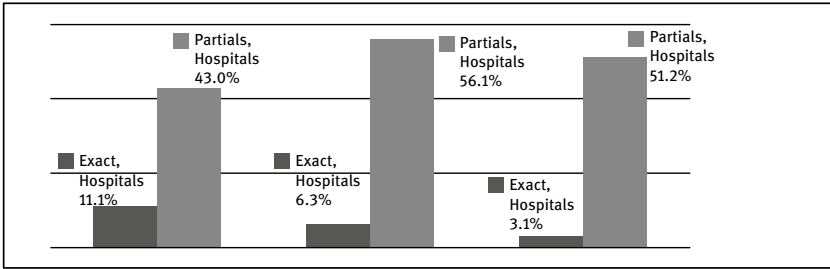


Fig. 10.8: Relative proportion of videos with MeSH terms.

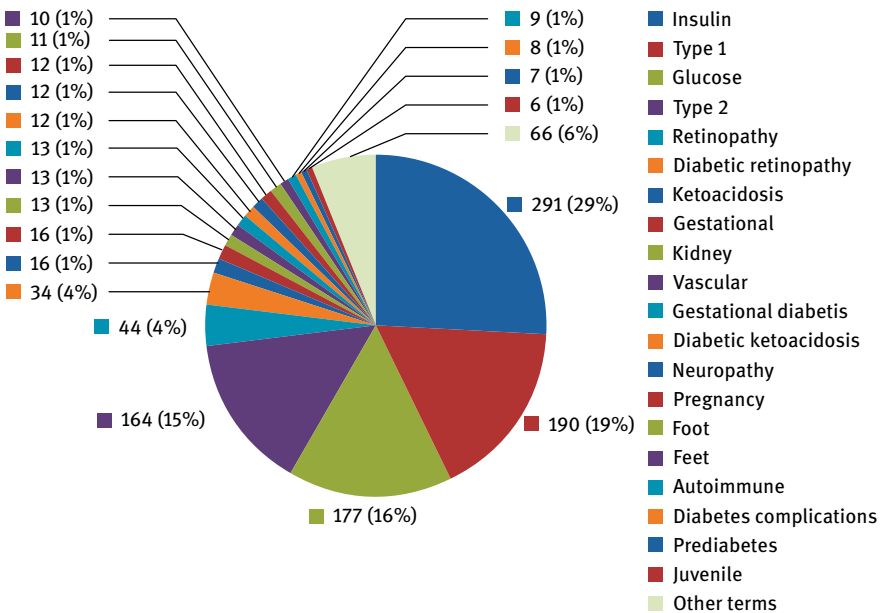


Fig. 10.9: The most frequently used MeSH terms in video title/descriptions.

represent 74% of the total amount of occurrences, while 41 terms are found among the remaining 26% of occurrences.

From the above analysis, we see that the MeSH results follow much of the same pattern as ICD-10 term detection. A main difference between ICD-10 and MeSH, is that a higher number of distinct terms were detected when using the MeSH vocabulary. We also found a higher number of exact match occurrences using MeSH.

10.3.3 Terms used in video titles and descriptions

The number of medical terms detected in video title and descriptions are few compared to the total number of diabetes related terms available from the ICD-10 and MeSH vocabularies. Among the 167 ICD-10 terms, only four exact terms were detected (2.4%), while 12 (out of 282) exact MeSH terms were detected (4.3%). Concerning partial terms, 23 were detected based on ICD-10 and 33 based on MeSH terms.

Below we present two tables listing the terms recognized in the exact term match (Tab. 10.2) and partial term match analysis (Tab. 10.3). Among the exact vocabulary terms from ICD-10 and MeSH, only 13 distinct terms were detected in video title and/or descriptions, while 36 distinct partial terms were found. Tables 10.2 and 10.3 list these terms, indicating to which vocabulary they belong, and provides the number of videos using each of these terms.

In addition to the partial terms listed in Tab. 10.3, we also detected 10 terms that appeared in title and/or description of one video only. These terms were: *acidosis, intolerance, nephropathy, glycosylation, fetal, pituitary, resistant, products, central, and sudden.*

We observe that a higher number of terms were detected when using the MeSH vocabulary. This may come as a consequence of the higher number of terms in the MeSH term set. Also, since the original use of MeSH is for indexing articles

Tab. 10.2: All exact terms detected in video title/descriptions. Including vocabulary affiliation and number of videos where the terms were used.

Exact terms	ICD-10	MeSH	Number of videos
Diabetic retinopathy	X	X	34
Gestational diabetes		X	13
Diabetic ketoacidosis	X	X	12
Insulin-dependent	X		8
Prediabetes		X	7
Diabetes complications		X	8
Diabetic diet		X	5
Diabetic neuropathy		X	5
Autoimmune diabetes		X	5
Diabetes mellitus	X	X	5
Mody		X	4
Diabetic coma		X	2
Glucose intolerance		X	1

Tab. 10.3: Partial terms detected in video title/descriptions. Including vocabulary affiliation and number of videos where the terms were used. Not including terms with only one occurrence.

Partial terms	ICD-10	MeSH	Number of videos
Insulin	X	X	291
Type 1	X	X	190
Glucose	X	X	177
Type 2	X	X	164
Retinopathy	X	X	44
Ketoacidosis	X	X	16
Gestational	X	X	16
Kidney		X	13
Vascular		X	13
Neuropathy	X	X	12
Pregnancy	X	X	12
Foot		X	11
Feet		X	10
Autoimmune		X	9
Juvenile		X	6
Dependent	X	X	5
Mellitus	X	X	5
Maturity		X	5
Hyperglycemic		X	4
Hyperosmolar	X	X	4
Renal	X		4
Peripheral	X		3
Syndrome	X	X	2
Abnormalities		X	2
Tolerance	X		2
Gastric		X	2

for PubMed, it might be natural (if one were to use medical terms) to choose a MeSH term also for describing videos.

From Tabs. 10.2 and 10.3 we see that the two vocabularies complement each other, since some of the detected terms are only found in ICD-10 while others only in MeSH. To check the overlap between the ICD-10 and MeSH term sets, we merged the two selected term sets. We found that by removing duplicates and plural forms of terms, the total amount of non repetitive terms were 344. Among these, the number of terms included in *both* ICD-10 and MeSH were 105. When we set aside the overlapped terms between these two thesauri, we see that both

ICD-10 and MeSH vocabularies contribute terms that are not identified in the other vocabulary respectively.

10.3.4 Occurrences of terms – when discarding the most common terms

As previously pointed out, the four partial terms (*type 1*, *type 2*, *insulin*, and *glucose*) had a very large number of occurrences when compared to the rest of the partial terms. To see how these four terms affect the total result, we made a new calculation of MeSH term matches, but now without recognizing the terms (*type 1*, *type 2*, *insulin*, *glucose*). We compared the new result (Fig. 10.10b) to the original result (Fig. 10.10a), and found that the number of “no match” videos dramatically increase (from 41% to 81%) when disregarding the (*type 1*, *type 2*, *insulin*, *glucose*) terms.

Figure 10.11 compares the relative proportion of partial MeSH term matches when all partial terms are included with those cases in which the four terms (*type 1*, *type 2*, *insulin*, *glucose*) were disregarded. When excluding the four most common terms, the proportion of videos including partial medical terms drops dramatically for videos from health organizations and active users, with a decrease of 83% and 87%, respectively. For hospital videos the proportion of partial MeSH term matches drops, but not to the same extent as for the two other groups. Here the decrease is 48%.

When disregarding the four most common terms, the distribution follows the same pattern as for exact term matches, where hospitals had the largest relative

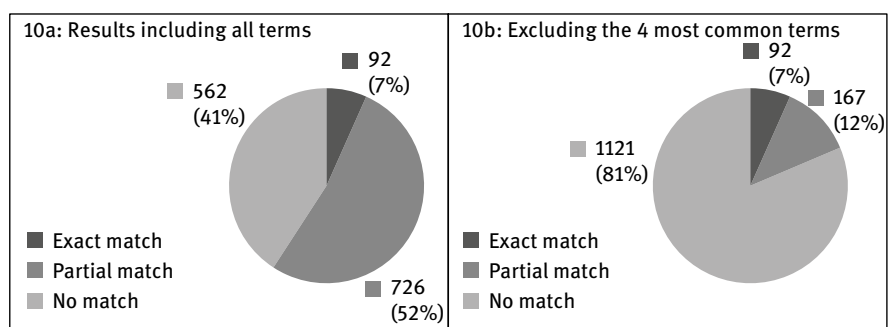


Fig. 10.10: The number of videos that have partial match for MeSH terms in title and/or description. In 10a all terms are included, while in 10b the four most common terms; (*type 1*, *type 2*, *insulin*, and *glucose*) were disregarded.

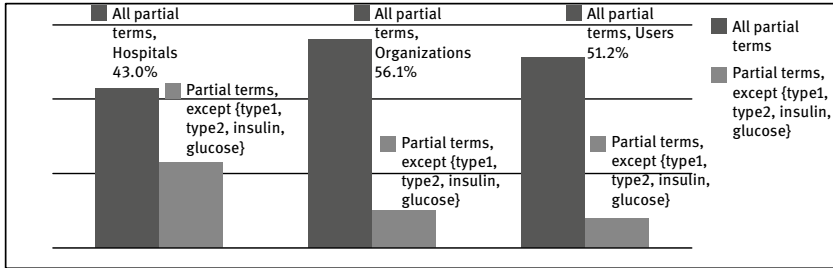


Fig. 10.11: Relative proportion of videos with MeSH partial terms. Analysis including all partial terms compared to analysis excluding the four most common terms (type 1, type 2, insulin, glucose).

proportion of videos including medical terms. This indicates that hospitals more often, compared to health organizations and active users, describe videos using the more specific medical terms.

10.4 Discussion

10.4.1 Findings

This study presents an analysis of medical terminology used in title and description of YouTube diabetes health videos uploaded by hospitals, health organizations and active users. The results show a low prevalence of medical terminology and that few distinct terms are actually used for describing videos.

We found that only 4% of the video title/descriptions included an exact ICD-10 term while 7% included an exact MeSH term. A larger amount of the total videos had a partial match with ICD-10 and/or MeSH terms. However, many of the positive results match one or more of the terms (*type 1, type 2, insulin, glucose*). When disregarding these very common terms, the amount of videos that include a partial MeSH terms was only 12%. The same applies to the findings of ICD-10 terms, where the amount of videos including a partial ICD-10 term drops to only 19% when disregarding the four common terms.

When distinguishing between videos provided by *hospitals, organizations, and active users*, we found that relative to the number of videos from each group, hospital videos had the highest prevalence of medical terms, followed by health organizations' videos and lastly by active users' videos. This was the case for exact term match to both ICD-10 and MeSH, and also for partial term match when disregarding the most common terms (*type 1, type 2, insulin, glucose*).

Even though hospital videos had the highest relative prevalence of medical terms (both exact and partial) the numbers were still low. Among hospital videos, only 11.1% included an exact MeSH term and 7.8% an exact ICD-10 term. The relative prevalence of partial MeSH terms (except the four common terms listed above) found in hospital videos was 22.2%.

In addition to low prevalence of medical terms found in video title and description, we also found a very low variety of diabetes-related medical terms in the title/description of such videos. Among the 167 ICD-10 terms, only four exact terms were detected (2.4%), while 12 (out of 282) exact MeSH terms were detected (4.3%). Concerning partial terms, 23 such terms were detected based on ICD-10 and 33 such terms were detected based on MeSH.

10.4.2 How ICD-10 and MeSH terms can be useful

Some hospitals and health institutions have for a long time provided patients with information about their disease and/or the medical procedure performed to help their ailment, especially on their discharge from treatment. This information normally contains very specific medical information that the patient may not only use to show to health personnel, explaining why they were at a hospital, but also to help them learn more about their health condition and how to prevent serious relapses.

In primary care, for example, there is an increasing trend towards providing patients with information about their condition. The medical terms present in the documentation given by primary care providers, especially the diagnostic or procedure codes, provide potential links to specific information that is relevant to the patient. For medical practitioners on the other hand such the diagnostic codes and MeSH terms can be used to identify literature that would be directly relevant for the patient through MEDLINE.⁵ The patient should also have the same opportunity to find relevant health information, including the wealth of medical information contained in patient health-related videos.

Providing both codes and the corresponding medical terms in video title and/or description, will help the patient to find information targeted at their medical condition. Codes and corresponding medical terms will also reduce the uncertainty that a patient will experience confronted with an often unknown and very complex medical reality. Being certain that you have identified information that

⁵ <http://www.nlm.nih.gov/bsd/pmresources.html>.

is relevant for the patients' medical condition can therefore be of great value for the patient.

Using medical terms and codes in the title and description of videos will also help producers of information resources, like YouTube videos, to identify gaps in information resources that patients may need. As the situation is now, neither the patient nor the producers of the videos are able to explore the collection in the same systematic manner as the medical professional explores the medical domain using MEDLINE.

10.4.3 Discriminating power of terms

The *specificity* of a term is a measure of its ability to distinguish between documents in a collection (Spärck Jones 1972; Salton & McGill 1983) or in our case, between health videos. It is common to measure a term's specificity using its Inverse Document Frequency (IDF) (Spärck Jones 1972; Baeza-Yates & Ribeiro-Neto 2011). The IDF measure is based on counting the number of documents, in the collection being searched, that contain the term in question. The intuition is that a term that occurs in many documents is not a good discriminator, and should be given less weight than one that occurs in few documents.

The results of our experiments show that the four partial terms (*type 1*, *type 2*, *insulin*, and *glucose*) have very low specificity among the diabetes related videos selected for this study. This reflects the common use of these terms. By describing a video using one or more of the most common terms, the video will be one among a large number of videos that are described in the same manner. To stand out, and be easy to identify and retrieve, a video needs a description that is (i) accurate with respect to video content and (ii) described through terms that have high specificity.

In this chapter we recommend providing textual descriptions to videos that also include medical terms and codes that as accurately as possible describe the content of the video. Following the consideration of specificity and discriminating power of terms, we also recommend avoiding the most common terms and rather describe video content through more specific medical terms.

10.4.4 The uniqueness of our study when compared to other work

A considerable amount of literature has been published on YouTube data analysis, such as studying the relations between video ratings and their comments

(Yee et al. 2009) or focusing on the social networking aspect of YouTube and their social features (Cheng, Dale & Liu 2008; Chelaru, Orellana-Rodriguez & Sengor Altingovde 2012). Up till now, studies of YouTube performance have mainly focused on YouTube in general rather than on specific domains, such as health. However, some studies have evaluated YouTube health video content with respect to their quality of information for patient education and professional training (Gabarron et al. 2013; Topps, Helmer & Ellaway 2013). Such studies, focusing on different areas of medicine, include the work of Butler et al. (2012), Briones et al. (2012), Schreiber et al. (2013), Singh, Singh & Singh (2012), Steinberg et al. (2010), Murugiah et al. (2011), Fat et al. (2011), and Azer et al. (2013). In these studies reviewers evaluate the quality of selected videos, and assess their usefulness as an information source within their respective area.

Here is an illustration of how our study differs from those of our colleagues. Konstantinidis et al. (2013) identified the use of SNOMED terms among tags attached to YouTube health videos. The videos, providing information about surgery, were collected from a preselected list of hospital channels. This study examined tags from 4307 YouTube videos and found that 22.5% of these tags was a SNOMED term. Our study, however, stands in contrast to the work done by Konstantinidis et al. in a number of ways:

First, we detect the use of terminology from *both* the ICD-10 and MeSH vocabularies in video title and descriptions, which we found to be interesting medical vocabularies because of their purpose and application area. The MeSH terms were originally used for indexing articles for PubMed. However, such terms might also be relevant when health video creators describe the content of their videos. ICD, which is the standard diagnostic tool for epidemiology, health management, and clinical purposes, is used to classify diseases and other health problems.⁶ As such, the ICD-10 vocabulary should also be relevant for describing health related content in videos. Second, in contrast to the work presented by Konstantinidis et al., we expanded the types of YouTube channels to also include health organizations and users active in publishing videos about the disease, as opposed to just looking at hospital-produced YouTube videos. Third, we focused on a narrowly-defined medical condition, examining a broad range of health videos containing information specifically about diabetes, as opposed to the work of Kostantinidis et al. who examined health videos on the more general topic of surgery. For all these reasons, both individually and collectively, we feel that we can provide a richer and fuller understanding of how educational health videos

⁶ <http://www.who.int/classifications/icd/en/>.

measure up, when looking at how title and description conform to the medical terms that pertain to the health condition for which the user is seeking crucial web-based video material.

10.5 Conclusion

We have investigated the prevalence of medical terminology in title and description of diabetes videos obtained from YouTube. Our results show that hospitals and health organizations only to a modest degree use medical terms to describe the diabetes health videos they upload to YouTube.

When comparing video title and description to terminology from the ICD-10 and MeSH vocabularies, we found that only 4% of the videos were described using an exact ICD-10 term and 7% using an exact MeSH term. Also, when disregarding four very common terms (i.e., *type 1*, *type 2*, *insulin*, and *glucose*), the amount of videos that included a partial MeSH term was only 12%. We further found that very few of ICD-10 and MeSH terms (2.4% and 4.3%, respectively) were used for describing diabetes health videos. This resulted in a very low variety of diabetes-related medical terms found in such videos.

We believe that including medical terms and codes when describing videos can improve the availability of the videos and make it easier for patients to find relevant information during an online search. Given the trend for hospitals, health institutions and primary care facilities to use medical terms and codes when providing patients with information about their conditions (which is done via the distribution of health-related pamphlets and newsletters as well as in email alerts sent out to patients), if health-related video descriptions would likewise include these important medical terms and related codes, the patient could then obtain videos relevant to their condition by simply using these terms and codes in their search queries.

When including medical terms in title and/or description of videos, one should also consider the specificity of terms, and their ability to discriminate between videos. One should avoid the most common terms (in our tests these were *type 1*, *type 2*, *insulin*, and *glucose*) and select, instead, more specific medical terms to describe this useful and informative video content so that such material would be more accessible to patients.

Acknowledgments

The authors appreciate support from UiT The Arctic University of Tromsø through funding from Tromsø Research Foundation, and the important contributions of the ITACA-TSB Universitat Politècnica de València group.

References

- AlGhamdi, K. M. & Moussa, N. A. (2012) 'Internet use by the public to search for health-related information', *Int J Med Inform*, 81(6):363–373, ISSN 1386-5056, doi: 10.1016/j.ijmedinf.2011.12.004.
- Azer, S. A., AlGrain, H. A., AlKheilaif, R. A. & AlEshaiwi, S. M. (2013) 'Evaluation of the educational value of youtube videos about physical examination of the cardiovascular and respiratory systems', *J Med Internet Res*, 15(11):e241.
- Baeza-Yates, R. A. & Ribeiro-Neto, B. (2011) *Modern Information Retrieval, the Concepts and Technology Behind Search*, Second edition. Bosto, MA: Addison-Wesley Longman Publishing Co., Inc.,
- Bennett, E. (2011) *Social Media and Hospitals: from Trendy to Essential*, *Futurescan Newsletter*, Healthcare Trends and Implications 2011–2016, The American Hospital Association, Health Administration Press.
- Bermudez-Tamayo, C., Alba-Ruiz, R., Jiménez-Pernett, J., García-Gutiérrez, J. F., Traver-Salcedo, V. & Yubraham-Sánchez, D. (2013) 'Use of social media by Spanish hospitals: perceptions, difficulties, and success factors', *J Telemed eHealth*, 19(2):137–145.
- Briones, R., Nan, X., Madden, K. & Waks, L. (2012) 'When vaccines go viral: an analysis of HPV vaccine coverage on YouTube', *Health Commun*, 27(5):478–485.
- Butler, D. P., Perry, F., Shah, Z. & Leon-Villapalos, J. (2012) 'The quality of video information on burn first aid available on YouTube', *Burns*, 2012 Dec 26:1.
- Chelaru, S., Orellana-Rodriguez, C. & Sengor Altinogvde, I. (2012) Can Social Features Help Learning to Rank YouTube Videos? In *Proceedings of The 13th International Conference on Web Information Systems Engineering (WISE 2012)*.
- Cheng, X., Dale, C. & Liu, J. (2008) Statistics and social networking of YouTube videos, In *Proceedings of the IEEE International Workshop on Quality of Service*.
- de Boer, M. J., Versteegen, G. J. & van Wijhe, M. (2007) 'Patients use of the Internet for pain-related medical information', *Patient Educ Couns*, 68(1):86–97.
- Dutta-Bergman, M. (2003) 'Trusted online sources of health information: differences in demographics, health beliefs, and health-information orientation', *J Med Internet Res*, [Online]. Available: <http://www.jmir.org/2003/3/e21/HTML>.
- Fat, M. J., Doja, A., Barrowman, N. & Sell, E. (2011) 'YouTube videos as a teaching tool and patient resource for infantile spasms', *J Child Neurol*, 26(7):804–809.
- Fox, S. (2011a) Peer-to-peer healthcare. *Report, PewInternet, California HealthCare Foundation*; Feb.
- Fox, S. (2011b) The social life of health information. *Report, PewInternet, California HealthCare Foundation*; May.
- Freeman, K. S. & Spyridakis, J. H. (2009) Effect of contact information on the credibility of online health information, *IEEE Transactions on Professional Communication*, 52(2), June.
- Gabarron, E., Fernandez-Luque, L., Armayones, M. & Lau, A. Y. (2013) 'Identifying measures used for assessing quality of YouTube videos with patient health information: a review of current literature', *Interact J Med Res*, 2(1):e6.
- Griffiths, F., Cave, J., Boarderman, F., Ren, J., Pawlikowska, T., Ball, R., Clarke, A. & Cohen, A. (2012) Social networks – The future for health care delivery, *Soc Sci Med*, 75(12), December, Available: <http://dx.doi.org/10.1016/j.socscimed.2012.08.023>.
- Karlsen, R., Borrás-Morell, J. E. B., Salcedo, V. T. & Luque, L. F. (2013) 'A domain-based approach for retrieving trustworthy health videos from YouTube', *Stud HealthTech Inform*, 192:1008.

- Konstantinidis, S., Luque, L., Bamidis, P. & Karlsen, R. (2013) 'The role of taxonomies in social media and the semantic web for health education,' *Methods Inf Med*, 52(2):168–179.
- Mishoe, S. C. (2008) 'Consumer health care information on the Internet: does the public benefit?,' *Respir Care*, 53(10):1285–1286.
- Moturu, S. T., Lui, H. & Johnson, W. G. (2008) Trust evaluation in health information on the World Wide Web, *Engineering in Medicine and Biology Society*, EMBS 2008. 30th Annual International Conference of the IEEE.
- Murugiah, K., Vallakati, A., Rajput, K., Sood, A. & Challa, N. R. (2011) 'YouTube as a source of information on cardiopulmonary resuscitation', *Resuscitation*, 82(3):332–334.
- NLM (2013) Fact Sheet Medical Subject Headings (MeSH®) [cited 2013 Dec 12]. Available from: <http://www.nlm.nih.gov/pubs/factsheets/mesh.html>.
- Powell, J., Inglis, N., Ronnie, J. & Large, S. (2011) The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study, *J Med Internet Res*, 13(1):e20. doi: 10.2196/jmir.1600. <http://www.jmir.org/2011/1/e20/v13i1e20>.
- Purcell, G. P., Wilson, P. & Delamothe, T. (2002) 'The quality of health information on the Internet', *Br Med J*, 324:557–8.
- Salton, G. & McGill, M. J. (1983) *Introduction to Modern Information Retrieval*, New York, NY: McGraw-Hill Inc.
- Schreiber, J. J., Warren, R. F., Hotchkiss, R. N. & Daluiski, A. (2013) 'An online video investigation into the mechanism of elbow dislocation', *J Hand Surg Am*, 38(3):488–494.
- Singh, A. G., Singh, S. & Singh, P. P. (2012) 'YouTube for information on rheumatoid arthritis – a wakeup call?' *J Rheumatol*, 39(5):899–903.
- Spärck Jones, K. (1972) 'A statistical interpretation of term specificity and its application in retrieval', *J Doc*, 28.
- Steinberg, P. L., Wason, S., Stern, J. M., Deters, L., Kowal, B. & Seigne, J. (2010) 'YouTube as source of prostate cancer information', *Urology*, 75(3):619–622.
- Syed-Abdul, S., et.al. (2013) 'Misleading health-related information promoted through video-based social media: anorexia on YouTube', *J Med Internet Res*, 15(2):e30.
- Topps, D., Helmer, J. & Ellaway, R. (2013) 'YouTube as a platform for publishing clinical skills training videos', *Acad Med*, 88(2):192–197.
- WHO (2013) International Classification of Diseases (ICD) Information Sheet, WHO. [cited 2013 Dec 12]. Available from: <http://www.who.int/classifications/icd/factsheet/en/index.html>.
- Yee, W. G., Yates, A., Liu, S. & Frieder, O. (2009) Are Web User Comments Useful for Search? In: *Proc. of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR09)*.

Editor's biography

Amy Neustein, Ph.D., is Editor-in-Chief of the *International Journal of Speech Technology* (Springer), a member of De Gruyter's STM Editorial Advisory Board, and Editor of their new series, *Speech Technology and Text Mining in Medicine and Health Care*. Dr. Neustein is also Series Editor of *SpringerBriefs in Speech Technology: Studies in Speech Signal Processing, Natural Language Understanding, and Machine Learning*. She has published over 40 scholarly articles, is a frequent invited speaker at natural language and speech technology conferences, and has given grand round lectures and seminars at over 20 leading medical institutions. She is editor of the volume, *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*. Dr. Neustein is co-editor of numerous books including *Forensic Speaker Recognition; Where Humans Meet Machines; and Mobile Speech and Advanced Natural Language Solutions*. She has been a member of the visiting faculty at the National Judicial College since 1985, and a member of MIR (Machine-Intelligence Research) Labs since 2010. She is the recipient of several distinguished awards: pro Humanitate Literary Award; Information Technology: New Generations (Medical Informatics) Award; and the Los Angeles County Supervisor Humanitarian Award. Dr. Neustein is Founder and CEO of Linguistic Technology Systems, located in Fort Lee, New Jersey.