Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing

Workshop Programme

31 May 2014

09:15 – 09:30 Welcome

09:30 – 10:30 Invited presentation: Georgios Paliouras, *The BioASQ Project*

10:30 – 11:00 Coffee break

11:00 – 11:20 Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith and Graciela Gonzalez, *Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark*

11:20 – 11:40 Takashi Okumura, Eiji Aramaki and Yuka Tateisi, *Expression of Laboratory Examination Results in Medical Literature*

11:40 – 12:00 Erwin Marsi, Pinar Öztürk, Elias Aamot, Gleb Sizov and Murat V. Ardelan, *Towards Text Mining in Climate Science: Extraction of Quantitative Variables and their Relations*

12:00 – 12:20 Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset and Pierre Zweigenbaum, *The Quaero French Medical Corpus: A Resource for Medical Entity Recognition and Normalization*

12:20 – 14:00 Lunch break

14:00 – 14:20 Rezarta Islamaj Doğan, W. John Wilbur and Donald C. Comeau, *BioC and Simplified Use of the PMC Open Access Dataset for Biomedical Text Mining*

14:20 – 14:40 Kirk Roberts, Kate Masterton, Marcelo Fiszman, Dina Demner-Fushman and Halil Kilicoglu, *Annotating Question Types for Consumer Health Questions*

14:40 – 15:10 Short Poster Presentations

Xiao Fu and Sophia Ananiadou, *Improving the Extraction of Clinical Concepts from Clinical Records*

Kerstin Denecke, *Extracting Medical Concepts from Medical Social Media with Clinical NLP tools: A Qualitative Study*

Mariana Neves, Konrad Herbst, Matthias Uflacker and Hasso Plattner, *Preliminary Evaluation of Passage Retrieval in Biomedical Multilingual Question Answering*

Borbála Siklósi, Attila Novák and Gábor Prószéky, *Resolving Abbreviations in Clinical Texts without Pre-Existing Structured Resources*

Lina Henriksen, Anders Johannsen, Bart Jongejan, Bente Maegaard and Jürgen Wedekind, *Worlds Apart - Ontological Knowledge in Question Answering for Patients*

Behrouz Bokharaeian, Alberto Díaz, Mariana Neves and Virginia Francisco, *Exploring Negation Annotations in the DrugDDI Corpus*

Stefan Schulz, Catalina Martinez Costa, Markus Kreuzthaler, Jose Antonio Miñarro-Giménez, Ulrich Andersen, Anders Boeck Jensen and Bente Maegaard, *Semantic Relation Discovery by using Co-occurrence Information*

Lorraine Goeuriot, Wendy Chapman, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling and Sanna Salanterä, *Building Realistic Potential Patients Queries for Medical Information Retrieval Evaluation*

15:10 – 16:30 Poster Session (Continuing into the Coffee Break)

16:00 – 16:30 Coffee break

16:30 – 18:00 Structured Discussion

- What are the most needed resources for health and biomedical text processing that currently don't exist or are not available?
- What would their availability enable us to do that we can't do now?
- Why are these resources not available or don't exist?
- How can we make them available or create them?

Editors

Sophia Ananiadou	University of Manchester, UK
Khalid Choukri	ELDA, France
Kevin Bretonnel Cohen	University of Colorado, USA
Dina Demner-Fushman	National Library of Medicine, USA
Jan Hajic	Charles University Prague, Czech Republic
Allan Hanbury	Technical University of Vienna, Austria
Gareth Jones	Dublin City University, Ireland
Henning Müller	HES-SO Valais, Switzerland
Pavel Pecina	Charles University Prague, Czech Republic
Paul Thompson	University of Manchester, UK

Workshop Organisers

Sophia Ananiadou	University of Manchester, UK
Khalid Choukri	ELDA, France
Kevin Bretonnel Cohen	University of Colorado, USA
Dina Demner-Fushman	National Library of Medicine, USA
Jan Hajic	Charles University Prague, Czech Republic
Allan Hanbury	Technical University of Vienna, Austria
Gareth Jones	Dublin City University, Ireland
Henning Müller	HES-SO Valais, Switzerland
Pavel Pecina	Charles University Prague, Czech Republic
Paul Thompson	University of Manchester, UK

Workshop Programme Committee

Olivier Bodenreider	National Library of Medicine, USA
Wendy Chapman,	University of Utah, USA
Hercules Dalianis	University of Stockholm, Sweden
Noémie Elhadad	Columbia University, USA
Graziela Gonzalez	Arizona University, USA
Jin-Dong Kim	DBCLS, Japan
Dimitris Kokkinakis	Gothenburg University, Sweden
Ioannis Korkontzelos	University of Manchester, UK
Hongfang Liu	Mayo Clinic, USA
Naoaki Okazaki	Tohoku University, Japan
Arzucan Özgür	Bogazici University, Turkey
Claire Nedellec	INRA, France
Sampo Pyysalo	University of Turku, Finland
Fabio Rinaldi	University of Zurich, Switzerland
Andrey Rzhetsky	University of Chicago, USA
Guergana Savova	Children's Hospital Boston and Harvard Medical School, USA
Hagit Shatkay	University of Delaware, USA

Rafal Rak	University of Manchester, UK
Lucy Vanderwende	Microsoft, USA
Karin Verspoor	NICTA, Australia
John Wilbur	NCBI, NLM, NIH, USA
Stephen Wu	Mayo Clinic, USA
Pierre Zweigenbaum	LIMSI, France

Table of contents

Author Index

Introduction

This volume contains the papers accepted at the *4rd Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014)*, held at LREC 2014, Reykjavik. Over the past years, there has been an exponential growth in the amount of biomedical and health information available in digital form. In addition to the 23 million references to biomedical literature currently available in PubMed, other sources of information are becoming more readily available. For example, there is a wealth of information available in clinical records, whilst the growing popularity of social media channels has resulted in the creation of various specialised groups. Extensive information is also available in languages other than English.

With such a deluge of information at their fingertips, domain experts and health professionals have an ever-increasing need for tools that can help them to isolate relevant nuggets of information in a timely and efficient manner, regardless of both information source and mother tongue. However, this goal presents many new challenges in analysis and search. For example, given the highly multilingual nature of available information, it is important that language barriers do not result in vital information being missed. In addition, different information sources cover varying topics and contain differing styles of language, while varying terminology may be used by lay persons, academics and health professionals. There is also often little standardisation amongst the extensive use of abbreviations found in medical and health-related text.

Applications in the health and biomedical domain are reliant on high quality resources. These include databases and ontologies (e.g., Biothesaurus, UMLS Metathesaurus) and lexica (e.g., BioLexicon and UMLS SPECIALIST lexicon). Given the frequently changing and variable nature of biomedical terminology and abbreviations, combined with the requirement to take multilingual information into account, there is an urgent need to investigate new ways of creating, updating such resources, or adapting them to new languages. New techniques may include combining semi-automatic methods, machine translation techniques, crowdsourcing or other collaborative efforts.

Community shared tasks and challenges (e.g. Biocreative I-IV, ACL BioNLP Shared Tasks (2009-2011-2013) etc.) have resulted in an increase in the number of annotated corpora, covering an ever-expanding range of sub-domains and annotation types. Such corpora are helping to steer research efforts to focus on open research problems, as well as encouraging the development of increasingly adaptable and wider coverage text mining tools. Interoperability and reuse are also vital considerations, as evidenced by efforts such as the BioCreative Interoperability Initiative (BioC) and the UIMA architecture. Several of the corpora introduced above are compliant with both BioC and UIMA, and are available within the U-Compare and Argo systems, which allow easy construction of NLP workflows and evaluation against gold standard corpora. There is also a need to consider how resources and techniques can facilitate easier access to relevant information that is written in a variety of different languages. For example, can existing techniques and resources used for machine translation, multilingual search and question answering in other domains be adapted simplify access to multilingual information in the biomedical and health domains?

The papers in this volume exemplify the diversity of research that is currently taking place, and explain how some of the challenges introduced above are being addressed. A number of research topics involving clinical corpora are represented, including the extraction of concepts (Fu and Ananiadou), resolution of abbreviations (Siklós et al.) and automatic generation of queries from medical reports as a means of enhancing patients' understanding of eHealth data (Goeuriot et al). Research into helping patients and consumers to obtain answers to medical queries is also the topic of a number of other papers, including the classification of question types as a means to determine the best strategy for answer selection (Roberts et al.), and the exploration of co-occurrence data in

UMLS to infer non-ontological semantic relations for the construction a knowledge base to support patient-centred question answering (Schulz et al.). Henriksen et al. describe a system operating on the Danish language that answers patients' queries by combining the use of formalised knowledge from a different medical resource (SNOMED CT) with text-to-text generation in the form of document summarisation and question generation. Also in the area of question answering, Neves et al. present a multilingual system for biomedical literature, operating on English, German and Spanish. A new corpus of biomedical documents in French annotated at the entity and concept level (Névéol et al.) further demonstrates the increasing efforts to facilitate the development of domain-specific systems in languages other than English. Okumur et al. explore the literature from a different perspective, i.e., as a source of information about laboratory results for use in building a disease knowledge base. Completing the set of papers concerning biomedical literature, Doğan et al. addresses the important issue of interoperability, through the description of the BioC annotation format, and the introduction of the BioC-PMC dataset, which contains annotations that conform to this description.

Exploiting the valuable knowledge available within the increasing volume of social media data is the subject of two papers. One explores the extraction of medical concepts from medical social media (Denecke), whilst the other describes a new corpus of tweets annotated with adverse drug reactions (Ginn et al.). The importance of studying the effects of drugs is also highlighted by Bokharaeian et al., who have enriched the DDI corpus, containing information about interactions between drugs, with negation cues and scopes. Finally, taking inspiration from research into biomedical literature, Marsi et al. explore the first steps in extracting entities and events from neighbouring research fields, i.e., climate science, marine science and environmental science, through the design of a new annotation scheme.

We wish to thank the authors for submitting papers for consideration, and the members of the programme committee for offering their time and effort to review the submissions. We would also like to thank our invited speaker, Dr. Georgios Paliouras, for his contribution.

Sophia Ananiadou, Khalid Choukri, Kevin Bretonnel Cohen, Dina Demner-Fushman, Jan Hajic, Allan Hanbury, Gareth Jones, Henning Müller, Pavel Pecina and Paul Thompson

Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark

Rachel Ginn[1], Pranoti Pimpalkhute[1], Azadeh Nikfarjam[1], MS, Apurv Patki[1], Karen O'Connor[1], Abeed Sarker[1], PhD, Karen Smith[2], PhD, Graciela Gonzalez[1], PhD

[1]Arizona State University, Tempe, AZ; [2]Regis University, Denver, CO

E-mail: rginn@asu.edu, ppimpalk@asu.edu, anikfarj@asu.edu, apatki1@asu.edu, Karen.Oconnor@asu.edu, abeed.sarker@asu.edu, ksmith003@regis.edu, Graciela.Gonzalez@asu.edu

Abstract

With many adults using social media to discuss health information, researchers have begun diving into this resource to monitor or detect health conditions on a population level. Twitter, specifically, has flourished to several hundred million users and could present a rich information source for the detection of serious medical conditions, like adverse drug reactions (ADRs). However, Twitter also presents unique challenges due to brevity, lack of structure, and informal language. We present a freely available, manually annotated corpus of 10,822 tweets, which can be used to train automated tools to mine Twitter for ADRs. We collected tweets utilizing drug names as keywords, but expanding them by applying an algorithm to generate misspelled versions of the drug names for maximum coverage. We annotated each tweet for the presence of a mention of an ADR, and for those that had one, annotated the mention (including span and UMLS IDs of the ADRs). Our inter-annotator agreement for the binary classification had a Kappa value of 0.69, which may be considered substantial (Viera & Garrett, 2005). We evaluated the utility of the corpus by training two classes of machine learning algorithms: Naïve Bayes and Support Vector Machines. The results we present validate the usefulness of the corpus for automated mining tasks. The classification corpus is available from http://diego.asu.edu/downloads.

Keywords: adverse drug reactions, twitter, social media, mining, machine learning, biomedicine, pharmacovigilance, classification, natural language processing

1. Introduction

Adverse drug reactions (ADRs), defined as "*injuries resulting from medical drug use*", present a significant health problem. A systematic review of twenty-five prospective observational studies determined that 5.3% of all hospital admissions are associated with adverse drug reactions (Kongkaew, Noyce, & Ashcroft, 2008). Accelerating the detection of these events could greatly impact human health.

To aid in ADR detection, many national reporting tools and social support networks have been developed. These can be accessed online by patients, practitioners, and researchers alike. Self-reported patient information, in particular, captures a valuable perspective that might not be captured in any other way. The information voluntarily submitted by patients to national agencies, like the US FDA's *MedWatch* program or the UK MHRA's *Yellow Card Scheme* is estimated to reflect less than 10% of the adverse effect occurrences (Inman & Pearce, 1993; Yang *et al.*, 2012). Thus, critical ADRs may go undetected until the harm done grows to a noticeable level.

Social media networks can present an alternative to formal national agency sites, and could prove to be a promising resource for ADR detection. Patients often submit pharmaceutical information in a wide variety of social networking resources, such as disease specific communities, blogs, microblogs, public news websites, or drug discussion forums. There has been a significant interest in these alternative sources for ADR detection in the last few years, starting with Leaman *et al.* (2010), which focused on comments from Daily Strength[1], a health-related community forum.

Twitter as a source of such comments, in general, presents different challenges. Here, we provide a manually annotated corpus from Twitter and outline these challenges and the methods we used to mine the Twitter microblogging platform for ADRs. The corpus contains 10,822 tweets annotated by domain experts for the presence of an ADR (as a binary attribute). In addition, it contains annotations for spans of text referring to specific ADRs along with UMLS IDs for the concepts.

Natural language processing from social media text is very challenging for any purpose, given that the text is highly unstructured and informal, and may contain a large number of misspelled words. For health-related mining, the challenges compound. Specific to our problem, identifying social media postings that reference a prescription drug is just the beginning of the problem; colloquialisms, hypothetical postulation, or information not relevant to personal experiences need to be filtered. For example, users might re-tweet news reports or comments about side effects heard on a television commercial. Even if the data were more structured and formal, automatic identification of ADRs is itself a challenging task. This is due to the complex relationships between drugs and their indications, adverse effects, and beneficial effects, as well as the great diversity of informal and creative terms that can match adverse-effect lexicon terms. In other words, if a patient simply mentions that a drug "*makes them sleepy*" it may apply to the drug's treatment effectiveness (as a

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number 1R01LM011176. The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health."

[1] http://www.dailystrength.org/

Research reported in this publication was supported by the

sleep aid), a beneficial effect (positive, but unintended effect), or an ADR (adverse effect). Furthermore, manually annotating data to structured vocabulary codes located in a dictionary (lexicon codes) increases complexity. Mentions like "*sleepy*", "*tired*", "*groggy*", and "*excessive sleepiness*" all have different lexicon codes (described in section 3.3) without normalization, and consequently, obscure the data and diminish automatic learning model accuracies.

We focus this paper on a description of the corpus and the process followed to monitor Twitter for comments related to a drug. The correct acquisition of tweets (taking into account misspellings, for example), adequate pre-processing, and systematic manual annotation are paramount to optimize the performance of machine learning methods trained using this data. We present a benchmark text mining application (classification of tweets as to whether they include an ADR mention or not) to demonstrate the potential utility of the corpus. Classification was performed using two classes of machine learning algorithms: Naïve Bayes (NB) and Support Vector Machines (SVM).

2. Related Work

In the past, a handful of different social media resources have been used in detecting ADR information. We can broadly group the efforts into those trying to exploit the data produced by online communities and disease-specific blogs, and those attempting to mine social media sites such as Twitter. The differences between them have not been well-studied, but data from online communities is generally more structured, since comments are usually grouped by treatment and/or disease. For example, two articles focused solely on data from a *diabetes* forum are discussed in the following subsection. We outline related work for each of these two broad categories, to better appreciate the differences.

2.1 Online Communities and Blogs

A few studies have explored the potential of the postings on the Daily Strength online community to detect ADRs. The work by Leaman *et al.* (2010), presented a lexicon based approach to detect ADRs. In addition, they compared the frequencies of ADRs in user comments to the frequency of documented ADRs. They explained that the most common sources of errors arise from novel/creative phrases, idiomatic expressions, string matching problems associated with misspellings, ambiguity, and miss-categorizations (identifying an indication as an ADR).

In order to address some of the limitations of the lexicon-based approach, Nikfarjam & Gonzalez (2011) proposed a method to capture the underlying syntactic and semantic patterns from the Daily Strength reviews. This study benefited from a large, manually annotated corpus (1,200 records) and an algorithm that can detect expressions not included in a lexicon. However, there are some limitations associated with the pattern-based method that prevent it from being a stand-alone solution for this task. One of the main challenges is its dependence on the size of the training data, since it identifies an extraction pattern only if enough matching sentences are observed in the training data

Two different studies were published in 2013 that utilized diabetes focused data. Akay, Dragomir & Erlandsson (2013) developed a methodology (text mining and self-organizing maps) to correlate positive and negative word cluster groups and with medical drugs and devices. This could aid the detection of ADRs through preprocessing for semantic tone/drug relationships prior to more complex ADR analysis. Likewise, Liu & Chen (2013) utilized the diabetes forum from Daily Strength (different than the treatment comments used by the above), and created a platform called AZDrugMiner that can be used on a variety of online blogs. This framework was evaluated with manually annotated forum posts and broken down into four different methodologies: medical entity extraction (a lexicon based approach utilizing MetaMap[2]), adverse drug event extraction (transductive SVM classifier on labeled an unlabeled data), report source classification (SVM classifier to differentiate personal experience from hearsay using labeled and unlabeled data), and analysis of ADR reports in the FDA Adverse Event Reporting System (FAERS) [3] *vs*. reports in patient forums (similarities and differences for various drugs).

Yang *et al*. (2012) analysed the MedHelp health community for drug safety signal detection. They relied on lexicon based ADR matching, association mining, and calculation of the proportional reporting ratio. Their study was limited to 10 predetermined drugs and 5 predetermined ADRs. Thus, they did not identify unknown ADRs without prior conditions — the authors simply looked for presence of pre-specified ADRs.

2.2 Twitter

Twitter, one of the largest social media websites, has over 645,000,000 users (as of January 1, 2014) and grows by an estimated 135,000 users every day, generating about 9,100 tweets every second[4] — a potential gold mine of information for researchers interested in studying population trends. This is especially true given Twitter's application programming interface (API), which makes part of its data publicly available and easily accessible. A recent survey revealed that 26% of online adults discussed personal health issues and that 42%[5] of them use social media to post or seek information about health conditions (Parker *et al.*, 2013). Tempting as it is to assume this is all easy to access and process, utilizing this enormous amount of succinct, unstructured, and informal information in a

[2] http://metamap.nlm.nih.gov/

[3] http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm

[4] http://www.statisticbrain.com/twitter-statistics/

[5] http://www.businesswire.com/news/home/20121120005872/en/Twenty-percent-online-adults-discuss-health-information#.UvQ4M4WmWGQ

way that can be useful to public health officials goes beyond the inherent challenges of natural language processing of such data, but automatic processing and extraction is a necessary first step. We outline here the advancements in the extraction of adverse drug reactions from twitter in the past two years.

Nakhasi *et al.* (2012) developed a manually annotated corpus of 770 tweets for patient safety event mentions that were preventable, adverse, care related, explicitly the result of health care actions/procedures, and something experienced on a first hand or someone personally known. The annotators noted who reported the event, the error source, error type, and error emotion. The results indicated that as much as 22.2% of the errors were medication related. However, this corpus was not intended to develop a natural language processing tool or algorithm to extract information — it was an annotated corpus and analysis of its promising statistical value.

Bian *et al.* (2012) created an SVM classification model from a large dataset — 2 billion tweets — on a high performance computing platform. The model sought to find prescription drug users and potential adverse events using 5 investigational cancer drugs. All data from a single user was aggregated into one document for temporal analysis by the SVMs — this ensured analysis of comments that spanned multiple tweets. Overall, 239 users were annotated for personal relationship to drug effects and 72 (the positive cases for personal relationships) were annotated for adverse events (resulting in 27 adverse events). Since the performance was limited, the authors suggested a number of error sources: noise (fragmented sentences, misspellings, non-word terms, odd-abbreviations, etc.), errors in tagging by MetaMap (Aronson & Lang, 2010), errors due to non-standard terms, and part-of-speech tagging errors due to highly unstructured text.

In another study (Jiang & Zheng, 2013), 5 drugs with established market presence were studied, collecting 6,829 tweets, and a classification model for drug effects was developed. Three classification models were analysed, with personal pronouns and sentiment analysis, to determine if tweets regarded "personal experiences." MetaMap was used to evaluate drug effects, which included diseases, findings, injuries, dysfunctions, and symptoms. The findings were compared to information on PatientsLikeMe[6] and MedLinePlus[7] with a 74-86% matching rate. However, the corpus studied was not annotated by domain experts and the study was not focused on ADRs.

Our corpus has broader coverage than the ones noted, with 74 carefully selected drugs queried, including misspellings, with a total of 10,822 tweets manually annotated by experts. It includes both binary annotation for classification applications, and specific concept annotation with mappings to UMLS concept IDs for the approximately 1,200 tweets that include a mention of an adverse reaction, indication, or beneficial effect. This can facilitate the development of advanced concept extraction and identification techniques for ADRs in Twitter.

3. Methods

Our manually annotated corpus was created through extraction of tweets related to 74 drugs of interest, using their brand and generic names, and phonetic misspellings. These were annotated for the presence of ADRs (a binary attribute for each tweet), the location/span of the reaction mentions, and the UMLS concept IDs for the ADR mentions. The overall process flow is shown in Figure 1 and subsequently described.

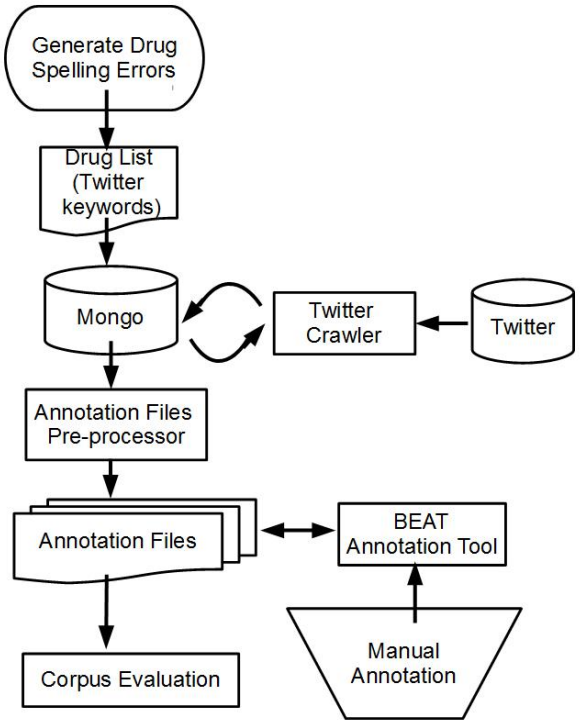


Figure 1: Data collection and annotation flowchart.

3.1 Drug List Generation

As part of a larger study, we selected a set of drugs to be monitored for different kinds of adverse effects. It includes both generally used drugs whose adverse effects are well known (*truth set*), and drugs released between 2007 and 2010 for which not all adverse effects are yet known and will only become visible as the drugs are more widely used. Drugs in the *truth set* were selected on the basis of their widespread use, as demonstrated by their presence in the Top 200 products by volume in the U.S. market. Many of the drugs for the *truth set* have come into widespread use recently, which allows for testing the capability of the natural language processing so at the time of release. For the newer drugs, going back to 2007 allows for market growth leading to common prescribing and comments on the site. The list was narrowed based upon forecasts for widespread use, the prevalence of disease states and conditions, and on whether the drug was new in class. Major categories

[6] http://www.patientslikeme.com/
[7] http://www.nlm.nih.gov/medlineplus/

include drugs for the central nervous system and neurological conditions such as Alzheimer's disease and schizophrenia. Treatments for age-related diseases like diabetes, cardiovascular diseases, urinary dysfunction, and musculoskeletal disorders also met the criteria for potential widespread use, given our life expectancy.

The list of drugs used as keywords to monitor Twitter was first expanded by including the generic and brand names of the drugs. Then, we extended the drug list to include misspelled drug names. This was critical to obtaining relevant tweets, as drug names are often misspelled in social media. We generated the misspellings through a 'phonetic spelling filter' (Pimalkhute *et al.*, 2013). This gave preference to variants that reflect the phonemes of the correct spelling. Examples of variants and misspelled tweets are shown in Table 1. Initially, the tool generated a large number of misspellings, out of which 18% were added to the list of drug names. This percentage was experimentally determined to maximize the tweet coverage while minimizing the number of terms needed to query Twitter. This is important because twitter API allows only 400 keywords per application key. This technique allowed us to capture an estimated 50 to 56% of tweets mentioning the drug.

Original Drug Name	**Example variants**
Prozac	prozaac, prozax, prozaxc,
Paxil	paxl, pxil, paxol.
Seroquel	seroquels, seroqul, seroqual
Olanzapine	olanzapin, olanzapoine, olanzaoine
Example Tweets with Seroquel Spelling Variants	
*@PsychoIogicaI HA! Not if you're on # **Seroquil** . EXTREMELY vivid dreams that stay in conscious memory. Very # Freaky ! Any idea why?*	
*@BipolarBlogger did you ever try the **Seriquel** XR??? It has a less sedative effect and has a longer lasting effect*	
*Gone from 50mg to 150mg of **Serequel** last night. Could barely wake up this morning and I feel like my body is made of lead*	
*@AndrewH_Smith Is the Inderal helpful? And yeah, they are short lasting but non addictive. You could try **Seraquel** too but it's pretty strong*	

Table 1: Examples of spelling variants that were generated and tweets that contained drug misspellings.

3.2 Preprocessing

After finalizing the drug name list, we accessed the twitter database through their publicly available API. The API allowed us to obtain matching tweets up to a volume equal to the streaming cap (~1% of all public tweets), restricted to 1000 requests per day[8]. This resulted in 187,450 tweets over a period of 6 months, which were then filtered to remove advertisements. To remove advertisements, we removed tweets that contained URLs. This cut down the tweets to 71,571.

Next, we balanced the dataset to select a set for annotation. This helped prevent dominance of some drugs over other drugs, as some drugs are much more popular than others. For example, 58,000 tweets were about nicotine; while other drugs averaged 240 tweets. We randomly selected a maximum of 300-500 tweets per drug, for a total of 10,822 tweets. All twitter datasets were stored and retrieved for later use using a Mongo[9] database for access by the annotators.

3.3 Manual Annotation

We sought to annotate our corpus not only for the presence or absence of ADR mentions, but also to identify the span of the expressions conveying individual ADRs, and to map them to formal medical terminology (*i.e.*, assigning them UMLS concept IDs). For the purpose of annotation, an ADR was defined as: "*an effect of the drug, which is not desired, and includes mentions that described a worsening of the patient's initial health condition*". It was important to distinguish these from *indications*, defined as: "*the causal condition, symptom, or disease that was the reason for the patient taking the drug*". Both these types of concepts are specifically annotated in the corpus. The distinction between ADRs and indications highlights the benefits of manual annotation as opposed to dictionary matching since the two types share concept names and UMLS codes, and would be indistinguishable if not by context.

The lexicon used to select the UMLS IDs was developed by augmenting a prior lexicon used in Leaman *et al.*, (2010). Originally, the lexicon included groups of terms from COSTART[10], SIDER Version I (Kuhn *et al.*, 2010), MedEffect[11], and a limited list of idiomatic expressions. The augmented lexicon used here also includes terms from SIDER Version II (Kuhn *et al.*, 2010) and the Consumer Health Vocabulary (CHV) (Zeng-Treitler *et al.*, 2008). We narrowed the scope of the terms in the UMLS to signs or symptoms, and excluded terms relating to topics like medical procedures. Overall, this resulted in 7,483 unique terms (UMLS concept IDs) and 16,182 concept names, which are normalized to the unique terms.

Next, two annotators manually annotated the processed tweets. The corpus comprises two types of data: binary annotation of ADRs and *full* ADR annotations (specified span and UMLS concept IDs). Chronologically, the annotators first analysed all 10,822 the tweets for the binary annotations. Following that, the tweets with ADRs were separated for full annotation. Both binary and full annotations were performed using an unpublished tool developed in-house.

[8] https://dev.twitter.com/discussions/4120

[9] https://www.mongodb.org/

[10] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/

[11] http://hc-sc.gc.ca/dhp-mps/medeff/index-eng.php

Annotators held weekly meetings to discuss the annotated tweets, correctness of concept labels, and develop annotation guidelines. The two annotators have medical or biological science background. Some meetings also included the full project team (one biomedical informatics student with a computer science background, two computer science students, and a pharmacology doctor). They annotated a total of 10,822 tweets, utilizing the following general principles for the concept annotation:

- Location boundaries of every mention should be minimized but the boundaries must also capture the entire concept
- Every annotation should be normalized to a UMLS concept ID that most *closely* matches the meaning
- For indirect matches, the most *general* ID should be used (may require annotator deliberation)

For instance, "*weight gain*", "*gained 20 pounds*", "*put on too much weight*", or "*fat fat fat*" would all be annotated to a general concept ID for "*weight gain*". Instances that caused confusion in selecting the most general term were discussed during meetings. For more information on the annotation process, please refer to the annotation guideline that accompanies the corpus.

3.4 Binary Classification

To demonstrate the utility of the corpus in detecting the tweets containing ADR mentions, we used NB and SVM classifiers for the binary classification task. Our intent was to investigate if supervised machine learning models can be trained on the annotated data in our corpus so that the tweets containing ADR mentions can be automatically identified. We used two classes for the experiments: *hasADR* and *noADR* — representing instances that contain ADR mentions and those that don't, respectively.

We performed preprocessing of the text to reduce noise. All of the words were lemmatized and transformed to lower case letters. Forms were also normalized. We implemented normalization through a dictionary algorithm that expands abbreviations (from a manually created list of commonly used abbreviations). For instance, "*abt*" was normalized to "*about*". We also implemented an algorithm similar to the one proposed by Brody & Diakopoulos (2011) for reducing word length in regards to emphasis words. For example, words like "*coollllllll*" or "*cooool*" were reduced to "*cool*".

For modeling text as vectors, we represented text instances as vectors in space of vocabulary (defined as set of all the unique words in the corpus). We used a simple term frequency scheme for vectorization, where value of i^{th} feature in the vector is equal to the number of times that feature or word occurs in that particular instance. For the SVM classifier, we used a linear kernel, and made no attempts at parameter optimization.

The binary dataset is very imbalanced and the number of noADR instances is much greater than the hasADR instances; thus, we trained the classifier on a more balanced dataset using randomly selected noADR instances. We created three sub-datasets varying in the skewedness towards the noADR class. The first dataset (dataset-1) is a balanced dataset, implying equal distribution of both classes: 1,008 hasADR, 1,008 noADR. The second dataset (dataset-2) has 60% noADR instances (1,008 hasADR, 1,512 noADR). The third dataset (dataset-3) has 70% noADR instances (1,008 hasADR, 2,352 noADR).

4. Results and Discussion

After developing the corpus, we analyzed our results to better understand its attributes and potential usefulness for text mining applications. We looked at the frequency and distribution of ADR mentions, the agreement between annotators (Table 2), and the performance on text mining classifiers.

4.1 Corpus Description and Statistics

A total of 10,822 tweets were annotated for the presence of ADRs by two experienced annotators, yielding approximately 1,200 tweets with at least one ADR mention (one annotator reported 1,008, while the other reported 1,255 tweets with 1,256 and 1,436 ADRs mentioned, respectively). For the experiments described in this paper, we considered the annotations by the first annotator to be the gold standard. In the final version of the corpus, the disagreements in the annotations will be resolved by Dr. Karen Smith.

To quantify the disagreements between annotators, we compared their annotations using partial matching criteria for span, and exact matching for concept IDs and the binary annotation. We report precision, recall and F-measure (Table 2) for agreement. For concept spans, an agreement (TP) occurs if there is some overlap on what the two annotators mark as the span for a concept. For concept IDs, annotators had to map an annotated concept to semantically equivalent UMLS concept IDs for an agreement to occur. For binary annotations, we computed the inter-annotator agreement (IAA) using Cohen's kappa (Carletta, 2006) and obtained a value of 0.69. According to Viera & Garrett (2005), a kappa of 0.61-0.80 indicates "substantial agreement".

As expected, concept IDs presented a larger source of disagreement than the spans of the ADR mentions within the text. However, we obtained fairly high values overall, especially for recall.

Comparison Type	Precision	Recall	F–Measure
Span	0.810	0.980	0.887
Concept ID	0.728	0.969	0.831
Binary	0.816	0.883	0.845

Table 2: IAA for concept span, concept ID, and binary (presence or absence of ADRs) annotations.

4.2 Binary Classification Results

To evaluate the performance of the classifiers, we computed evaluation metrics (precision, recall, f-measure and accuracy) using 10-fold cross validation. Table 3 shows the performance of the NB and SVM classifiers for binary classification. The equations for the metrics are shown in Equations 1-4.

	noADR			**has**ADR			
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Accuracy**
Dataset – 1 (balanced)							
NB	0.608	0.840	0.705	0.891	0.695	0.781	0.746
SVM	0.799	0.684	0.737	0.625	0.754	0.683	0.715
Dataset – 2 (60%)							
NB	0.670	0.877	0.760	0.852	0.638	0.730	0.746
SVM	0.845	0.739	0.789	0.456	0.654	0.538	0.728
Dataset – 3 (70%)							
NB	0.745	0.883	0.808	0.759	0.563	0.646	0.752
SVM	0.894	0.796	0.842	0.445	0.653	0.529	0.766

Table 3: Classification results for the NB and SVM classifiers on the three data sets.

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP})}$$

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FN})}$$

$$\boldsymbol{Fmeasure} = \frac{\boldsymbol{2 * precision * recall}}{(\boldsymbol{precision} + \boldsymbol{recall})}$$

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{Total\ correct\ predictions}}{\boldsymbol{Number\ of\ test\ instances}}$$

Equations 1-4: Formulaic equations for the classifier evaluation metrics.

4.3 Error Analysis

4.3.1. Corpus and Annotation

The disagreements in annotations between the annotators were analysed for the three categories: concept ID disagreement, span differences and nonmatching annotations. The largest source of disagreement came from the differences in concept IDs selected. There are several key reasons behind these discrepancies. The first is the similarity, and in some cases exact matches, in terms found in the lexicon associated with different concept ID numbers. For example, a patient may mention a "*hangover*" as an ADR. In the lexicon there is a concept ID for "*pill hangover*" and another ID for "*hangover from alcohol or any other drug substance*", the similarities of the terms make it difficult to ensure agreement as there is no obvious reason to select one over the other. Another source of discrepancies in concept ID annotations comes from differences in the annotators' interpretation of the idioms, slang or euphemisms used by the patient. The relatively small size of each tweet exacerbates this problem because it can eliminate the clues to meaning that can often be found in the context of large text documents.

The span and nonmatching annotation discrepancies mostly arise from the same issue, *i.e.*, differing interpretations between annotators regarding how much of the text should annotated to capture the concept. This can be problematic at two levels: not only does it lower IAA, but the span selected by an annotator tends to influence the concept ID selected, further adding to disagreements. For example, annotating "*tremors in hands*" *vs*. just "*tremors*" resulted not only in span disagreement but also in concept disagreement with the first being mapped to the ID for "*tremor of hands*" and the latter to "*tremors, shaking*". Nonmatching annotations are instances where one annotator annotates a portion of the text that was not annotated at all by the other annotator. Disagreements of this nature often highlight the difficulties annotators can have in determining whether the person is discussing their own adverse experience with the drug or if they are merely providing a commentary on the drug. One such case is the following tweet: '*depression hurts, cymbalta can help? one of that s*** many awful symptoms is thoughts of suicide..*.'. One annotator selected "*thoughts of suicide*" as an ADR and the other did not, based on the criterion that an ADR was to be annotated only if it was

experienced by the user posting the mention. The first annotator interpreted the message as such, while the second did not. Once again, the limited nature of tweets removes contextual clues and annotations become a subjective decision made by the annotators. Future work to improve IAA might include the review and revision of the lexicon to unify concepts that are the same or similar. To improve issues with span selection, ongoing annotation meetings to discuss and revise the annotation guidelines with the purpose of improving clarity and conciseness of instruction should help annotator decisions in the future.

4.3.2. Binary Classification

To obtain estimates about the relative performances of the NB and SVM classifiers on this data, we compared their accuracies against the *majority labeling* baseline. Majority labeling is the method of labeling every test instance to the majority class in the test set. This evaluation enabled us to (i) compare the performances of the classifiers against a simple baseline that does not use lexical data from the corpus, and (ii) compare the performances of the classifiers for balanced *vs.* unbalanced data sets. Figure 2 illustrates the results.

From the Figure, it can be observed that both classifiers perform better than the majority labeling baseline, even when the data set is heavily imbalanced (*i.e.*, the majority class represents 70% of the data). This clearly indicates that the data and annotations in our corpus aid the training of automatic models for classification of ADRs. It can also be observed that the classifier accuracies increase slightly as the proportion of the majority class increases. Our inspection of the results show that this is because as the number of instances for the majority class increases, so does the accuracy over the majority class. Since the overall accuracy relies more on the majority class as its proportion increases, increase in the classification accuracy for the majority class results in overall accuracy increase as well. This, however, does not mean that the accuracy for the hasADR class increases as well as the imbalance in the data increases. The overall effect of the training set proportions on the classification accuracy for the hasADR class requires further experimentation. In this analysis, we did not attempt to deeply analyze how the training and test set proportions affect classification accuracy, or what training-test ratio would be ideal for the classification of real-life data. We leave this as future work.

We analyzed the false positives associated with the hasADR. In many of the instances, the author would talk positively about a drug, but our supervised learning algorithms, based on the presence of specific terms, classify the statements as ADR mentions. For example: "*my [drug] is kicking in, I can feel it.*" Such errors may perhaps be eliminated by using more intelligent feature selection techniques for the classifiers. A number of false positives contain mentions from the ADR lexicon, but the statements do not actually report adverse effects. Instead they present the authors' questions (*e.g.*, "*who've been prescribed [drug]) for sleep? Has it helped at all?*"), or just life stories (*e.g.*, "*I could pass out with a moments notice even with this [drug] in my system*"). Another group of errors are tweets with idiomatic expressions and sarcasms, such as "*I took a [drug] and yet I\'m in an awesome mood. This never happens*".

Many of the misclassifications are due to the lack of deep semantic analysis of the lexical contents of the tweets. However, such techniques are beyond the scope of this paper as our intent is to present the corpus and investigate its potential for the implementation of advanced automatic techniques for ADR detection. The performances of our classifiers using basic lexical features are very promising. We leave the implementation of more complex techniques for classification as future work.

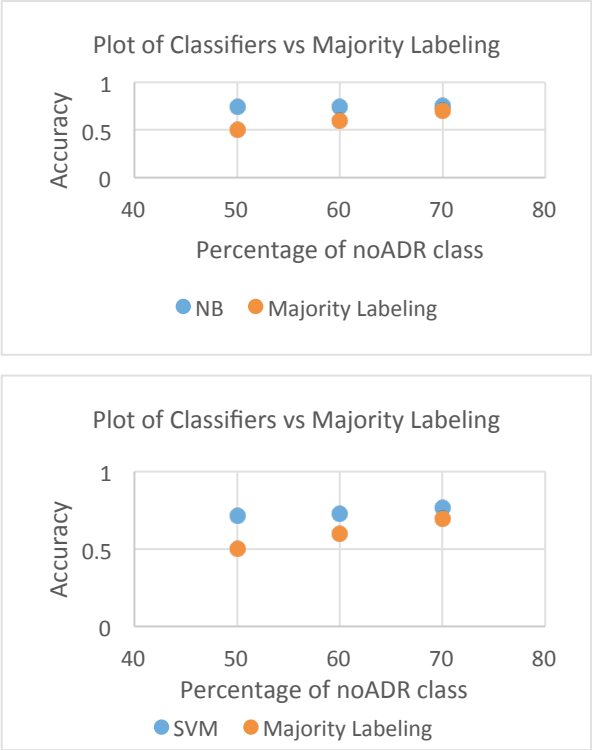


Figure 2: Plots of the classifiers *vs.* majority labeling: Naïve Bayes on top, SVM at the bottom. Both converge as the training data becomes highly unbalanced.

5. Conclusions

We presented an annotated Twitter corpus focused on ADR mentions with broad pharmacological coverage, collection twitter comments about 76 drugs. Only 65 of the drugs had one or more associated tweets, with wide variability in the number of tweets per drug. We selected a balanced number of tweets per drug, to form a corpus that contains a total of 10,822 tweets manually annotated by experts. It includes both binary annotation for classification applications, and specific concept annotation with mappings to UMLS concept IDs for the

approximately 1,200 tweets that included a mention of an adverse reaction, indication, or beneficial effect. This can facilitate the development of advanced concept extraction and identification techniques for adverse drug reactions in Twitter. The binary annotations are available for download at http://diego.asu.edu/downloads We show the utility of our corpus by applying two supervised machine learning approaches for the binary classification task of identifying if tweets contain ADR mentions or not. Although the classifier performances are modest, it is intended as a baseline for future development. Multi-stage natural language processing platforms could be applied for the binary classification and other associated ADR detection tasks. We applied the NB and SVM classifiers using surface level lexical features (*i.e.,* word vectors). Our goal is to incorporate features through the use of deep semantic analysis of the text associated with the tweets.

References

Akay, A., Dragomir, A., & Erlandsson, B. E. (2013). A novel data-mining approach leveraging social media to monitor and respond to outcomes of diabetes drugs and treatment. In *Point-of-Care Healthcare Technologies (PHT), 2013 IEEE* (pp. 264–266). IEEE.

Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229–236.

Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. In: *Proceedings of the 2012 international workshop on Smart health and wellbeing*. ACM; 2012:25–32.

Brody S, Diakopoulos N. Cooooooooooooooollllllllll-l!!!!: Using Word Lengthening to Detect Sentiment in Microblogs. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011:562–570.

Carletta J. Assessing agreement on classification tasks: the kappa statistic. *Comput Linguist*. 1996;22(2):249–254.

Inman, W., & Pearce, G. (1993). Prescriber profile and post-marketing surveillance. *The Lancet*, *342*(8872), 658–661.

Jiang, K., & Zheng, Y. (2013). Mining Twitter Data for Potential Drug Effects. In *Advanced Data Mining and Applications* (pp. 434–443). Springer.

Kongkaew, C., Noyce, P. R., & Ashcroft, D. M. (2008). Hospital admissions associated with adverse drug reactions: a systematic review of prospective observational studies. *The Annals of Pharmacotherapy*, *42*(7), 1017–1025.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., & Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular Systems Biology*, *6*(1).

Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, 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* (pp. 117–125). Association for Computational Linguistics.

Liu, X., & Chen, H. (2013). AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *Smart Health* (pp. 134–150). Springer.

Nakhasi A, Passarella R, Bell SG, Paul MJ, Dredze M, Pronovost P. Malpractice and malcontent: Analyzing medical complaints in twitter. In: *2012 AAAI Fall Symposium Series*.; 2012. 2014.

Nikfarjam, A., & Gonzalez, G. H. (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. In *AMIA Annual Symposium Proceedings* (Vol. 2011, p. 1019). American Medical Informatics Association.

Parker, J., Wei, Y., Yates, A., Frieder, O., & Goharian, N. (2013). A framework for detecting public health trends with Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 556–563). ACM.

Pimalkhute, P., Patki, A., Nikfarjam, A., & Gonzalez, G. (2013). Phonetic Spelling Filter for Keyword Selection in Drug Mention Mining from Social Media. In *AMIA SUMMIT* (Vol. 2014). American Medical Informatics Association.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med*, *37*(5), 360–363.

Yang, C. C., Jiang, L., Yang, H., & Tang, X. (2012). Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media. In *Proceedings of ACM SIGKDD Workshop on Health Informatics (August 12, 2012)*.

Zeng-Treitler, Q., Goryachev, S., Tse, T., Keselman, A., & Boxwala, A. (2008). Estimating consumer familiarity with health terminology: a context-based approach. *Journal of the American Medical Informatics Association*, *15*(3), 349–356.

EXPRESSION OF LABORATORY EXAMINATION RESULTS IN MEDICAL LITERATURE

Takashi Okumura, Eiji Aramaki, Yuka Tateisi

National Institute of Public Health, Kyoto University, National Institute of Informatics
Japan
taka@niph.go.jp, eiji.aramaki@design.kyoto-u.ac.jp, yucca@nii.ac.jp

Abstract

Medical literature contains expressions of laboratory examination results, which are invaluable knowledge sources for building a disease knowledge base that covers even rare diseases. In this study, we analyzed such expressions of disease descriptions in open databases with manually built dictionaries and obtained the following results. First, we identified two major types of expressions for laboratory examination results, that with and without their test names in the expressions. Second, the study identified evaluative expressions that frequently appear in the description of the results. Third, presence of test names and evaluative expressions could classify the expressions into four major classes that demand independent strategies to interpret. The study illustrated that this is beyond the scope of the existing corpora in this domain mostly designed for medical records. Although the analysis is based on rudimentary statistics, it clarified the factors necessary for future corpus designs to promote further research.
Keywords: Laboratory examination, Medical literature, Corpus design

1. Introduction

Clinical Decision Support Systems necessitate disease knowledge bases that describe the relationship between disease and their associated signs and symptoms, as well as typical laboratory results. Because manual compilation of such a knowledge base requires considerable efforts, using Natural Language Processing (NLP) is a reasonable approach. In this regard, medical NLP tools, such as MetaMap (Aronson, 2001), cTakes (Savova et al., 2010), and MedLEE (Friedman et al., 1995), should contribute to automating the compilation process to some extent by extracting the expression of **clinical findings** in the descriptions of target diseases.

On the other hand, the interpretations of **laboratory examination results**, particularly those of rare diseases, are available only in review articles and case reports, and a disease knowledge base demands the processing of expression of laboratory examination results in such medical literature. However, studies have mainly targeted the extraction of clinical findings from a large number of clinical documents, and only a few studies (Zhang and Patrick, 2006) have attempted to interpret expression of laboratory findings in medical literature, which are readily available in a machine readable format on hospital information systems nowadays (McDonald et al., 2003).

Accordingly, this paper analyzes expressions of laboratory examination results to clarify the factors for future corpus design to facilitate further research activities. To this end, Section 2. outlines laboratory examinations and their result expressions, and presents a hypothetical classification model with four major categories. For verification of the model, Section 3. develops simple dictionaries and reports on the statistics of the public disease databases OMIM (John Hopkins University, 1987) and Orphanet (INSERM SC11, 1997). Section 4. discusses the results and the considerations required for future corpus design. Section 5. reviews previous studies that are highly relevant to laboratory examination results and Section 6. concludes the paper.

2. Expression of laboratory results

2.1. Overview of laboratory examinations

Laboratory examinations are performed at medical institutions and involve analyzing samples collected from patients, with the aim of measuring some property of the sample. For example, the white blood cell (WBC) count is an examination performed using a blood sample to determine the number of WBCs in a unit volume. Laboratory examinations are categorized according to the samples used, such as blood, urine, and stool. They are also categorized according to the method of analysis, such as hematological, immunological, physiological, and microscopical means, each of which has certain characteristic style for their result descriptions. To illustrate the characteristics of their results, the examinations are classified by the type of results (Table 1), analyzing a catalog of laboratory examinations available in Japanese clinical settings (Fumimaro Takaku (ed.), 2009). As shown, examinations with numerical results were dominant in the catalog (40.8%), and expression of their results is represented simply using numbers or adjectives such as *high*, *normal*, and *low*. The second most abundant in the ranking is examinations yielding only high/low results, without numbers (24.7%), followed by those with positive/negative results (19.9%). The remaining examinations, such as microscopical examinations and blood typing yield qualitative results.

Result Category	# of records	
Numerical	334	(40.8%)
High/Low	202	(24.7%)
Positive/Negative	163	(19.9%)
Load/Function test	29	(3.5%)
Typing/Genetic test	21	(2.6%)
Fraction/Isozymes	18	(2.2%)
Microscopical findings	14	(1.7%)
Others	37	(4.5%)
Total	818	(100.0%)

Table 1: Classification of examination type

		Evaluative expression	
		+	−
Examination name	+	Quantitative examinations (E1,2,3,4,5,6,7)	Qualitative examinations
	−	Descriptive expression (E8,9,10,11,12)	Nominal expression (E13)

Table 2: Classification of the expressions for examination results

2.2. Expression of laboratory examination results

Laboratory examination results in medical literature are mostly appear in free format. To analyze the expressions, we randomly selected 20 entries from OMIM (John Hopkins University, 1987) that included "syndrome" in the title and had 450 as the last three digits of the record (Okumura et al., 2013), and found 13 expressions of laboratory examination results in the annotated descriptions, which were classified as follows.

Results with examination name The simplest form of laboratory result expression found was that with an examination name, which can be formalized as either "evaluative adjective + examination" or "examination + copula + adjective", as listed below. Note that characteristic adjectives, such as *elevated, decreased*, and *normal*, were found in the expressions.

1. *elevated mean pulmonary artery pressure*
2. *elevated liver function enzymes*
3. *elevated transaminases*
4. *moderately elevated alkaline phosphatase*
5. *normal esophageal manometry findings*
6. *SGPT was chronically elevated*
7. *hCG stimulation tests showed normal testosterone responses*

Results without explicit name The second category of result expressions are those without explicit examination names. For example, the results of loading tests can be expressed in a descriptive manner, without their examination names. The names are omitted in such expressions, probably because the findings and the responses are more essential than their formal names in the context and domain experts can easily infer the actual name if needed. Note that, in the examples below, evaluative adjectives are found, such as *low, effective*, and *insensitive*.

8. *deletion of part of the short arm of chromosome 5*
9. *lymphocyte infiltration of the CNS*
10. *low to low-normal gonadotropin (152760) responses*
11. *Administration of testosterone enanthate was effective*
12. *insensitive to growth hormone*

Nominal expression of examination results Lastly, we found unique expression for a laboratory result that does not contain an examination name, evaluative adjective, nor qualitative results.

13. *hypercholesterolemia*

Hypercholesterolemia implies that there was a sampling of blood from the patient and the cholesterol level was high in the biochemical examination result. In this expression, the name of the laboratory examination and its result are naturally encoded into the nominal form. Although this is an established expression of a certain clinical finding, it is a possible form of examination result that must be detected for disease knowledge bases that include laboratory examination results, because clinicians may perceive it also as increased cholesterol. Such expressions include *hyperkalemia, hyperglycemia, viremia, bacteremia, acidemia, acidosis, alkalosis*, and *alkalemia*, all of which correspond to certain examination results.

2.3. Classification of laboratory examination results and technical challenges

As illustrated, there are various ways of expressing laboratory examination results, and notably, there are expressions of examination results without the name of the examination. This observation suggests that expressions of laboratory examination results can be classified into two major categories, that with and without examination names. Further, some of the expressions contained evaluative adjectives and others did not, a distinction that can also be utilized for classification. The identified axes can categorize the expression into four types, as illustrated in Table 2. "En" in the cells refers to the 13 expressions.

Unfortunately, the examples lack an expression with an examination name that depicts results in a qualitative manner. However, there are various cases in clinical medicine that fall in this category; for example, result expressions for urinary sediment analysis contain various *casts* with independent names. More importantly, the table suggests the existence of expressions without examination names. An illustration is a descriptive expression, *insensitive to growth hormone*, which suggests a negative examination result of the growth hormone stimulation test. The expression describes actual examination in an operational manner, presupposing scientific domain knowledge on the part of the intended readers. Although the substance name can be used to restore the name in this case, there are examinations with *proper name*, where no such hint is available in the result expression. For example, "fragility of red blood cells" for a patient of paroxysmal nocturnal hemoglobinuria (PNH)

Normal	Abnormal	Increase	Decrease	Positive	Negative	Sufficiency	Deficiency
normal	abnormal	increase	decrease	positive	negative	sufficiency	deficiency
sensitive	insensitive	increased	decreased	above	below	sufficient	deficient
effective	ineffective	increasing	decreasing	high	low	excess	insufficiency
	defective	elevated	declined	higher	lower	excesses	insufficient
		elevation	decline			excessive	
		elevating	declining				
		rise	lowering				
		risen	diminished				
		rising	diminishing				
			reduced				
3 items	4 items	9 items	10 items	4 items	4 items	5 items	4 items

Table 3: Simple dictionary for evaluative expressions

suggests a positive result for the *Ham test*. Because clinicians may perceive it in either way, there is a challenge for medical NLP to recognize the expressions of laboratory examination results as such.

3. Statistical analysis of result expressions

The 2x2 table is a hypothetical model, inductively built only with a limited number of samples. However, to the best of our knowledge, there exists no corpus that annotates expression of laboratory examination results with appropriate information to validate the model. Accordingly, as a preliminary study, this section attempts to estimate the frequency of each category, by building simple dictionaries for expressions of laboratory examination results. Because each type of expression in Table 2 requires an independent strategy to interpret them, the breakdown of the expressions would also contribute to guide future research efforts.

3.1. Dictionary for evaluative expressions

First, to extract expressions of laboratory examination results, a list of evaluative adjectives found in such expressions was built. To this end, we identified eight categories and 43 items in total, as shown in Table 3. Because expressions of examination results may be expressed in various forms, such as a noun phrase ("*elevation of transaminases*") or as a sentence ("*transaminases were elevated*"), a stemmer may be used in preprocessing for more efficient matching. However, for the simplicity of statistics, the terms were manually nominalized, verbalized, and paraphrased in advance.

3.2. Dictionary for examination names

Development of a dictionary for laboratory examination names may seem easy, as it can be done simply by extracting the examination names from an examination catalog, or a standardized terminology (International Health Terminology Standard Development Organisation, 1999). However, such a dictionary would be of little practical use for the detection of the expressions.

For example, an examination, called "human leukocyte antigen typing", is frequently used in medical literature to denote a type of disorder with genetic background. However, it appears as "HLA typing" and "HLA findings". It is possible to avoid the complication by adding the acronym, HLA, to the dictionary. However, because there

Result Category	# of records	# of synonyms	
Numerical	334	532	(1.6)
High/Low	202	368	(1.8)
Positive/Negative	163	391	(2.4)
Load/Function test	29	52	(1.8)
Typing/Genetic test	21	32	(1.5)
Fraction/Isozymes	18	40	(2.2)
Microscopical findings	14	22	(1.6)
Others	37	60	(1.6)
Total	818	1497	(1.8)

Table 4: Overview of the examination name dictionary

are ambiguous acronyms, such as CSF, which is indicative of colony stimulating factor (CSF) or cerebrospinal fluid (CSF), such a simple strategy may not work as expected.

Another example is "creatine kinase isozymes" used to differentiate certain diseases, for which an examination to measure their levels is frequently used in clinical settings. However, there are a variety of descriptions found in the OMIM database (John Hopkins University, 1987) that refer to the result of the common examination. In the example below, BB, MB, and MM are the isozymes (types) of creatine kinase (CK).

- *The dimeric creatine kinase isozymes are involved in maintaining intracellular ATP levels, particularly in tissues that have high energy demands.*
- *The creatine kinase MM isozyme is found exclusively in striated muscle; the BB isozyme is found in smooth muscle, brain, and nerve; CKMB is found in human heart.*
- *demonstrated marked elevation of BB isozyme fraction of serum creatine kinase for male sibs with this disorder.*

Another type of difficulty arises from terminological variation. There is an examination that measures serum thyroid stimulating hormone (TSH) level, but another examination that measures "TSH receptor antibody" is also described as "antibody of TSH receptor", or "antibodies of TSH receptors", in actual literature. This circumstance also compromises simple dictionary matching strategies such as the longest match.

	OMIM		Orphanet	
Sentences with Examination Name (EN+)	13,343	(6.9%)	2,555	(7.8%)
Sentences with Evaluative Expression (EE+)	30,577	(15.8%)	4,566	(13.9%)
Sentences with both EN+ and EE+	4,742	(2.4%)	755	(2.3%)
Sentences with nominal laboratory results	4,836	(2.5%)	1,165	(3.6%)
Total # of sentences	193,687	(100.0%)	32,768	(100.0%)

Table 5: Number of sentences with examination names and evaluative expressions

		Evaluative expression		
		+	−	
Examination name	+	**2.3–2.4%**	4.5–5.5%	**6.9–7.8%**
	−	11.6–13.4%	**2.5–3.6%**	15.2–15.9%
		13.9–15.8%	7.0–9.1%	

Table 6: Number of sentences in the categories

As illustrated, automated generation of the dictionary is not a practical approach. Instead, we manually paraphrased examination names in the catalog for laboratory examinations (Fumimaro Takaku (ed.), 2009) and developed a dictionary by adding their synonyms and acronyms in succession, searching against the target document with the aim of better detection power. In this process, the catalog was used to extract laboratory examinations, simply because it is widely used in Japan and covers various categories of examinations for clinical purposes. Statistics of the resulting dictionary is shown in Table 4, with the number of synonyms compiled.

3.3. Dictionary for nominal laboratory result

As shown in Section 2., we found nominal expressions of examination results in medical literature. Such an expression can be tentatively defined as the expression of a finding or a medical state that is conclusively determined by a single laboratory examination. For example, *hyperkalemia* is identified by a test measuring serum potassium level, and thus such an expression is equivalent to "increased potassium". This circumstance suggests that the class of expressions may have characteristic prefixes, such as *hyper-* and *hypo-*, as well as a suffix, such as *-mia*.

To enumerate such expressions, the following operation was performed. First, the SNOMED CT vocabulary (International Health Terminology Standard Development Organisation, 1999) was used to extract terms that have the prefixes (*hyper-* and *hypo-*) and the suffix (*-mia*), resulting in 145, 102, and 476 terms, respectively. Second, SNOMED CT contains 2,285 concepts that have the label, "On examination", in their names, and 201 single-word concepts were selected for further processing. The resulting 924 terms contained findings that are simply discovered by physical examination, such as *hypodontia* (congenital absence of teeth), and disorders not diagnosed by a single laboratory finding alone, such as *ischemia*. Accordingly, the list was manually inspected to extract terms that denote clinical states determined only by a single laboratory test in ordinary clinical settings, and finally 244 terms were identified that matched the criteria.

3.4. Statistics and Analysis

Finally, items in the dictionaries were searched against the descriptions of diseases in public disease databases. To this end, the descriptions of diseases contained in OMIM (John Hopkins University, 1987) were downloaded, and then the GENIA Sentence Splitter (Sætre et al., 2007) was applied to the descriptions, resulting in 587,601 sentences, after the removal of duplications and predefined headings. Then, for 6,727 valid disease records, 193,687 sentences were selected for statistics, without further error correction. Preprocessing for Orphanet (INSERM SC11, 1997) was performed in a similar manner. Orphanet contained 6,442 disease records, but there were 3,073 records with an identical "under construction" message. The remaining 3,369 records were selected for further processing, and the sentence splitter yielded 32,768 sentences. Finally, items in the dictionaries were searched against the sentences, simply with the following command.

```
grep -i -w -f dictfile sentencefile
```

The result is shown in Table 5. Sentences with examination names accounted for 6.9–7.8% of the total sentences, and sentences with evaluative expression accounted for twice as many, accounting for 13.9–15.8% of the entire sentences. Sentences that contained both examination names and evaluative expressions accounted for 2.3–2.4%. In addition, nominal terms were searched in OMIM and Orphanet, and the expressions appeared in 2.5% and 3.6% of the sentences, respectively.

These numbers in Table 5 were then transformed into a 2x2 table (Table 6), for comparison with Table 2. In the table, bold items are measured in the analysis, and other items are calculated from the measured numbers. Table 6 suggests that there is a gap between the number of expressions that include (6.9–7.8%) and do not include (15.2–15.9%) examination names. A breakdown of the numbers indicates a further gap between evaluative expressions that contain (2.3–2.4%) and do not contain (11.6–13.4%) examination names.

The dictionary for evaluative expressions includes prepositions such as *above* and *below* that appear often in ordinary English writings, and adjectives such as *normal* and *abnormal* may be used for clinical findings, as well as for laboratory examination results. Accordingly, it is likely that statistics overestimates the number of evaluative expressions for laboratory result expressions (11.6–13.4%). In contrast, the number of expressions with neither examination names nor evaluative expressions (2.5–3.6%) accounted for only limited vocabulary, and the actual figure was underestimated. For example, medical literature occasionally contains acronyms such as *B27* and *Rh(+)*, which suggest that laboratory examinations for HLA typing and blood typing were performed. Because comprehensive listing of such expressions requires far greater costs, these cases are not included in this study, resulting in the underestimation. Combined together, although the simple statistics is not sufficiently accurate to draw a firm conclusion, the number (15.2–15.9%) suggests undocumented classes of expressions that merit further investigation.

4. Discussion

Although the estimation in the last section was rudimentary, the result suggested that there are expressions of laboratory examination results with and without their names in the expressions. This section discusses the findings of the preliminary analysis, in the light of future corpus design to interpret expressions of laboratory examination results.

First and foremost, the newer corpus has to address the omission of examination names, identified in the expressions of laboratory examination results. Existing annotation schemes might recognize examination names and results in expressions independently, and their relationship would be identified (Roberts et al., 2007). However, such a strategy can cover expressions that have simple relationship between examination names and their results. Accordingly, in the future corpus, several factors must be taken into consideration. Examinations with quantitative results can be expressed in a descriptive manner, with evaluative adjectives and without explicit reference to the examination name. In this class of expressions, the most complex case is exemplified in the *Ham test* example, where the restoration of the examination name requires scientific knowledge and inference. Examinations with qualitative results can also be expressed without their formal examination name; for example, those with microscopical identification. We also identified a nominalized form of laboratory result expression, such as *hypercholesterolemia*. For recognition of these expressions, they must be annotated as such, preferably associated with their restored examination names.

Second, we analyzed evaluative expressions (Table 3) in medical literature and proposed eight categories for evaluative expressions that represent certain aspects of such expressions. However, in clinical settings, "slight increase" and "severe increase" may have different meanings. For example, a "slight increase" of WBC may be observed even in healthy person, whereas a "severe increase" certainly suggests abnormality in the body. Accordingly, the *degree of change* must be taken into account in the annotation scheme.

Third, expressions of qualitative results also demand special handling, for example, results for various typing tests and genetic analysis. The nominal expressions of laboratory results, and the *Ham test* example might be included here. In this area, paraphrasing technologies (Androutsopoulos and Malakasiotis, 2010) may help to convert the expressions into more tractable form, given a sufficient number of annotated documents.

Lastly, complexity in the expressions of laboratory expressions must be addressed, in addition to the processing of names and evaluative expressions. This preliminary study revealed that only one-third of sentences with examination names contained an evaluative expression, suggesting that sentences displaying a simple relation between an examination name and an examination result are in the minority. Some sentences included multiple evaluative expressions and conjunctions, further complicating the interpretation process. Besides, the result expressions might deliver generalized information, as well as description of cases. The corpus needs to provide necessary information to appropriately handle such cases.

5. Related Work

The processing of expressions for laboratory examination results in free-text format has been studied in a very limited context. We performed systematic survey on the PubMed (National Library of Medicine, 1996) and the ACL Anthology (The Association for Computational Linguistics, 2002), and found just a few items in this problem domain.

Zhang (2006) suggested a tag, *medical test*, for the expressions of examination results in their corpus of clinical case reports. The paper randomly selected case reports in a database and removed reports that described groups of patients with the aim of extracting descriptions of individual patients. However, the report was not accompanied by in-depth analysis of the expressions, and the corpus is not available for research use. Okumura et al. (2013) presented a small corpus with similar tags for laboratory examination results in medical literature, and suggested the need for knowledge extraction of laboratory examination results, also without detailed analysis.

To recognize expressions of laboratory results, corpora for medical records may be reused, such as the CLEF corpus (Roberts et al., 2007) and the i2b2 corpus (Uzuner et al., 2011). However, as suggested throughout this paper, there is an unignorable amount of expressions for laboratory test results, that do not contain an explicit reference to examination names. Accordingly, existing corpora are unusable for knowledge acquisition of that type of results. Kang and Kayaalp (2013) shares the same limitations, which attempted to recognize four elements, (name of specimens, analytes, units of measures and detection limits) that explicitly appear in descriptions of laboratory tests. An exception is a study conducted by Bhatia et al. (2010) that attempted to interpret examination results for the management of diabetes mellitus patients, including weight, blood pressure, low-density lipoprotein (LDL), high-density lipoprotein (HDL), creatinine, total cholesterol (TC), glucose, fasting glucose, and glycated hemoglobin (HbA1C) levels, in medical records. The study reported 80.0–98.3% recall and

88.8–100% precision, which is impressive. However, descriptions for weight, blood pressure, and HbA1C levels are simple enough to be processed only with predefined rules, and the remaining tests contain very limited variations in their expressions. Accordingly, they are not applicable to a more general expression.

The interpretation of examination results appear in guidebooks for laboratory result interpretation (Wallach, 2007; Pagana and Pagana, 2009), but only common diseases are mostly covered, and a comprehensive knowledge source even for rare diseases is not yet available. Okumura and Tateisi (2013) analyzed the logical structure of such a guidebook and investigated the efficient acquisition and representation of the knowledge in the book. There are ontologies for laboratory examinations and their results, which were developed for the integration and exchange of laboratory data across institutions, but they are not designed for the interpretation of the results described in free format. (Baorto et al., 1997; Baorto et al., 1998; McDonald et al., 2003; Khan et al., 2006).

6. Conclusion

The relationship between laboratory examination results and their causes are available in guidebooks for laboratory examinations. However, the causes listed in the literature are common diseases, and rare diseases are seldom covered. Accordingly, the development of a disease knowledge base with broad coverage necessitates the processing of expressions for laboratory examination results, available in various medical literatures such as case reports and review articles. However, a detailed study of such expressions has not been explored in the field of medical NLP.

This study identified that expressions of laboratory examination results can be classified into two major categories, that with and without examination names. The expressions without examination names may seem counterintuitive, but it is natural for domain experts to omit formal names if there is more essential information than examination names in the context and if the description carries the information necessary to infer the names. The processing of such expressions requires scientific knowledge of laboratory examinations and an inference mechanism.

This study also identified evaluative expressions that tend to appear in quantitative expressions of laboratory examination results. The expressions are used for differentiating quantitative and qualitative results, each of which requires an independent strategy to interpret expression of laboratory examination results. This axis, coupled with the examination name axis, should define four classes of expression of laboratory examination results that must be identified to extract the necessary information.

Knowledge about the causal relationship between diseases and their laboratory findings is a key component of clinical decision support systems, and automated acquisition of this knowledge would contribute to their improved quality. However, existing corpora in medical NLP studies have been mostly designed to interpret clinical findings in medical records. Although the methodology used in this article is still preliminary and the data volume is limited, the results suggest that the interpretation of laboratory results expression in medical literature requires unique considerations that have not yet been covered by already published studies. To promote further research, it would be desirable to develop an appropriate corpus designed for the interpretation of medical literature.

7. References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38(1):135–187, May.

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *AMIA Annual Symposium*, pages 17–21.

David M. Baorto, James J. Cimino, Curtis A. Parvin, and Michael G. Kahn. 1997. Using logical observation identifier names and codes (loinc) to exchange laboratory data among three academic hospitals. In *Proceedings of the AMIA Annual Fall Symposium*, pages 96–100. American Medical Informatics Association.

David M. Baorto, James J. Cimino, Curtis A. Parvin, and Michael G. Kahn. 1998. Combining laboratory data sets from multiple institutions using the logical observation identifier names and codes (loinc). *International journal of medical informatics*, 51(1):29–37.

Ramanjot S. Bhatia, Amber Graystone, Ross A. Davies, Susan McClinton, Jason Morin, and Richard F. Davies. 2010. Extracting information for generating a diabetes report card from free text in physicians notes. In *the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 8–14. Association for Computational Linguistics.

Carol Friedman, Stephen B Johnson, Bruce Forman, and Justin Starren. 1995. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 347. American Medical Informatics Association.

Fumimaro Takaku (ed.). 2009. *Laboratory examinations databook 2009-2010*. Igaku-Shoin. (in Japanese).

INSERM SC11. 1997. Orphanet. http://www.orpha.net/.

International Health Terminology Standard Development Organisation. 1999. SNOMED CT. http://snomed.org/.

John Hopkins University. 1987. OMIM: Online Mendelian Inheritance in Man. http://www.ncbi.nlm.nih.gov/omim.

Yanna Shen Kang and Mehmet Kayaalp. 2013. Extracting laboratory test information from biomedical text. *Journal of pathology informatics*, 4.

Agha N. Khan, Stanley P. Griffith, Catherine Moore, Dorothy Russell, Arnulfo C. Rosario Jr, and Jeanne Bertolli. 2006. Standardizing laboratory data by mapping to loinc. *Journal of the American Medical Informatics Association*, 13(3):353–355.

Clement J. McDonald, Stanley M. Huff, Jeffrey G. Suico, Gilbert Hill, Dennis Leavelle, Raymond Aller, Arden Forrey, Kathy Mercer, Georges DeMoor, John Hook, et al. 2003. Loinc, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4):624–633.

National Library of Medicine. 1996. PubMed. http://www.ncbi.nlm.nih.gov/pubmed.

Takashi Okumura and Yuka Tateisi. 2013. Interpretation of laboratory examination results and their simple representation. In *IEEE International Symposium on Computer-Based Medical Systems (CBMS2013)*, June.

Takashi Okumura, Eiji Aramaki, and Yuka Tateisi. 2013. Classification and characterization of clinical finding expressions in medical literature. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM2013)*, December.

Kathleen D. Pagana and Timothy J. Pagana. 2009. *Mosby's manual of diagnostic and laboratory tests*. Elsevier Health Sciences.

Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Subbarao Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, et al. 2007. The CLEF corpus: semantic annotation of clinical text. In *AMIA Annual Symposium Proceedings*, volume 2007, pages 625–629. American Medical Informatics Association.

Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Y Matsubyashi, and Tomoko Ohta. 2007. Akane system: protein-protein interaction pairs in the biocreative2 challenge, ppi-ips subtask. In *Proceedings of the Second BioCreative Challenge Workshop*, pages 209–212.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

The Association for Computational Linguistics. 2002. ACL Anthology. http://aclweb.org/anthology/.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Jacques B. Wallach. 2007. *Interpretation of diagnostic tests*. Lippincott Williams & Wilkins.

Yitao Zhang and Jon Patrick. 2006. Extracting patient clinical profiles from case reports. In *the 2006 Australasian Language Technology Workshop (ALTW2006)*, pages 167–168.

Towards Text Mining in Climate Science: Extraction of Quantitative Variables and their Relations

Erwin Marsi[1], Pinar Öztürk[1], Elias Aamot[1], Gleb Sizov[1], Murat V. Ardelan[2]

[1]Department of Computer and Information Science, [2]Department of Chemistry
Norwegian University of Science and Technology (NTNU)
{emarsi,pinar,sizov}@idi.ntnu.no, eliasaa@stud.ntnu.no, murat.v.ardelan@ntnu.no

Abstract

This paper addresses text mining in the cross-disciplinary fields of climate science, marine science and environmental science. It is motivated by the desire for literature-based knowledge discovery from scientific publications. The particular goal is to automatically extract relations between quantitative variables from raw text. This results in rules of the form "If variable X increases, than variable Y decreases". As a first step in this direction, an annotation scheme is proposed to capture the events of interest – those of change, cause, correlation and feedback – and the entities involved in them, quantitative variables. Its purpose is to serve as an intermediary step in the process of rule extraction. It is shown that the desired rules can indeed be automatically extracted from annotated text. A number of open challenges are discussed, including automatic annotation, normalisation of variables, reasoning with rules in combination with domain knowledge and the need for meta-knowledge regarding context of use.

Keywords: Text Mining, Literature-based Discovery, Climate Science, Marine Science, Environmental Science, Corpus Annotation, Relation Extraction, Event Extraction

1. Introduction

One of the characteristics of the cross-disciplinary fields of climate science, marine science and environmental science is the existence of many different processes that affect each other in direct and indirect ways, resulting in highly complex systems. In climate science, for example, *climate feedback* is defined as "An interaction in which a perturbation in one climate quantity causes a change in a second, and the change in the second quantity ultimately leads to an additional change in the first. A negative feedback is one in which the initial perturbation is weakened by the changes it causes; a positive feedback is one in which the initial perturbation is enhanced." (Stocker et al., 2013). Identifying such feedback processes is generally considered a crucial step in understanding and predicting phenomena like global warming.

Unfortunately much potential knowledge regarding processes and their interactions is hidden in the scientific literature, scattered over journals catering to different scientific communities with relatively little communication among them. Given the vast and constantly growing nature of this body of literature, it is indeed hard for individual researchers to keep track of all relevant publications in their field of expertise, let alone of those in related or even more distant areas.

Text mining of scientific literature may contribute to alleviating this problem (Etzioni, 2011). Climate, marine and environmental science may all benefit from automatic extraction of processes and their interactions from scientific publications. Extracted information can be indexed, thus allowing researchers to search for interactions between processes in a much more effective way than with conventional keyword-based search engines.

In addition, this structured data can be used for inference in discovery support systems. For example, pairs of cause and effect processes can be chained together, possibly in combination with existing domain knowledge, in order to suggest hypotheses about indirect interactions or feedback loops or to point out contradictory findings. Such discovery of implicit knowledge in a body of literature is aimed for in the field of *literature-based discovery* (LBD). The first results in LBD were produced by Swanson (1986a) through manually executing a search algorithm based on co-occurrence statistics of terms. This allowed him to combine two publicly available knowledge fragments – (1) Fish oils reduce blood viscosity and (2) patients with Raynaud's disease tend to exhibit high blood viscosity – to form the hypothesis that fish oils treat Raynaud's disease. The hypothesis was later confirmed experimentally. Even though both knowledge fragments were publicly available to any researcher, nobody had been aware of both knowledge fragments and made the connection.

The aim of LBD is to create systems that provide discovery support to uncover such potential hypotheses, which Swanson referred to as *undiscovered public knowledge* (Swanson, 1986b). Most LBD methods are based on chaining of unspecified relations, using term co-occurrence frequencies as heuristic evidence for a relation between two terms. Terms are usually extracted as n-grams from the text (Lindsay and Gordon, 1999) or taken from a controlled vocabulary or ontology (Weeber et al., 2001).

Recently, co-occurrence based LBD methods have come under critique for yielding imprecise results, as they fail to exploit the true breadth of knowledge contained in the scientific literature. Hristovski et al. (2008) therefore advocate a text mining based approach, where relation extraction is used to discover specific relations between two concepts. This enables more precise and complex query patterns.

LBD efforts along these lines in the biomedical domain have taken advantage of existing tools such as SemRep

(Rindflesh and Fiszman, 2003). SemRep is a major text mining system for the biomedical domain that exploits structured domain knowledge found in the Unified Medical Language System (UMLS)[1]. The UMLS consists of three tools: a lexicon, a semantic network and a meta-thesaurus. Using underspecified symbolic language processing, SemRep is able to extract a wide range of specific relations, such as TREATS, HAS_PART and LOCATION_OF. However, adapting such tools to less resourced domains, such as marine science, is difficult because of the lack of resources like UMLS and because of the knowledge, time and effort required for writing extraction rules.

Other work has concentrated on relation and event extraction through machine learning. Causal relations are one of the most commonly targeted relations. In (Mihăilă and Ananiadou, 2013), several machine learning algorithms are applied to recognise causality triggers such as *therefore*, *because*, *as a result of*, etc. The approach is tested on BioCause (Mihaila et al., 2013) and BioDRB (Prasad et al., 2011) corpora, which consist of articles with manually annotated causal relations between named entities in the biomedical domain. A machine learning approach has also been applied by Pechsiri and Piriyakul (2010) to extract causal relations in the agricultural domain, which are then used to construct explanation knowledge graphs that represent the domain knowledge. Supervised learning requires domain-specific training material though, which is currently lacking in our domain of climate science, marine science and environmental science.

An alternative approach may be the use of unsupervised learning and clustering techniques. As an example of unsupervised techniques for causal relation extraction, Hashimoto et al. (2012) propose a set of relations that can be used to detect causality. They identify excitatory, inhibitory and neutral relations with a corresponding set of extraction templates. More templates are acquired automatically by a bootstrapping process. Excitatory relations are then used for extraction of contradictions, causality relations and generation of causality hypotheses. A disadvantage of these approaches is that their performance is generally far less accurate than that of supervised methods. In addition, it seems they are not capable of covering the more complex events of changing processes and their interactions, as we are interested in here.

Other approaches have explored extraction of more fine-grained types of events, including those of increase and decrease. Zambach and Lassen (2010) identify and linguistically analyse verbs that express regulation relations, positive and negative, between processes and substances in the biomedical domain. They suggest that their analysis can benefit extraction of, as well as reasoning over, these relations in the biomedical domain, although no implementation or evaluation was carried out.

In sum, currently text mining in climate science, marine science and environmental science appears to be virtually non-existent. Our research agenda targets developing text mining in this area, in particular towards applications in LBD. Existing approaches and tools from other domains such as biomedicine are not readily applicable and do not provide extraction of the type of processes and interactions needed for our purposes. Development of domain-specific event extraction tools is therefore high our agenda. Following the lead of text mining initiatives in biomedicine (Kim et al., 2009), we explore manual text annotation for creating annotated corpora, which can be used to train classifiers for automatic annotation, and ultimately automatic rule extraction. In Section 2., an annotation scheme is proposed to annotate the events of interest – those of change, cause, correlation and feedback – as well as the entities involved in them. Its purpose is to serve as an intermediary step in the process of rule extraction. It is shown in Section 3. that such rules can indeed be automatically extracted from annotated text. Section 4. discusses a range of open challenges, including automatic annotation, normalisation of variables, reasoning with rules in combination with domain knowledge and the need for meta-knowledge regarding context of use. The final Section 5. lists conclusions and future work.

2. Annotation

2.1. Procedure

The data consisted of 12 abstracts (2369 words) from recent, high-quality scientific journal publications about the relation between climate and ocean changes. These were selected by our domain expert (author MVA) as a reasonably representative sample of the text type in the targeted area, comprising multi-disciplinary work in marine biology, marine chemistry, oceanography, environmental science, climate science, biogeoscience and geophysics. Text was automatically extracted from PDF files. Abstracts were manually extracted, tokenised and split into sentences, also allowing for manual correction of minor PDF-to-text conversion errors.

Annotation was carried out using the Brat annotation tool (Stenetorp et al., 2012). The annotation scheme described below was developed in an iterative fashion in close collaboration with our domain expert. It is inspired by annotation efforts in the biomedical domain such as the GENIA corpus (Kim et al., 2003) and the corpora used in the BioNLP shared tasks on event extraction (Kim et al., 2009). It covers a particular type of events – those of change, cause, correlation and feedback – and the entities involved in them, quantitative variables. The primary reason for annotation is not to analyse the text according to some linguistic formalism or theory, or to follow some knowledge representation formalism or ontological theory. Instead the purpose of the annotation is rather pragmatic: to serve as an intermediary step in the process of extracting rules about the relation between quantitative variables from raw text.

2.2. Annotation scheme

The resulting annotation scheme involves one type of entity (variable), several types of events (change, increase, decrease, cause, correlate, feedback) and some basic logic structure (and/or, negation).

2.2.1. Variables

A quantitative variable is an entity that can be counted or measured. Its value can be naturally expressed by a number

[1] http://www.nlm.nih.gov/research/umls/

such as a count, a ratio, a percentage or a scalar (quantity of units). It can be regarded as a (potential) quantitative variable in an experiment or a model. Not every variable in the text is labeled as such. To save annotation time and effort, only those variables related to a change are annotated. The direction of change can be positive (increasing), negative (decreasing) or unspecified (either increasing or decreasing), but there must always be a clear cue in the text that the variable is involved in some change. Examples of changing, increasing and decreasing variables respectively:

(1) a. significant changes in [*surface ocean pH*]
b. rise in [*atmospheric CO2 levels*]
c. decline in [*marine primary production*][2]

In contrast, the text spans in (2) are not annotated as variables.

(2) a. *[*carbon dioxide*] and [*light*] are two major prerequisites of photosynthesis
b. *changes in [*the network of global biogeochemical cycles*]
c. *The concentrations of [*DFe*] and [*TaLFe*] were relatively high

The text spans in (2-a) are measurable, in principle at least, but there is no textual cue in the context indicating that they are subject to change. The text span in (2-b) admittedly identifies something that is changing, but it is an abstraction – not something that can be measured and naturally expressed through a number. The ones in (2-c) express a static state rather than a dynamic event. The reason for excluding cases like these is that they do not lead to useful rules about the relation between quantitative variables.
Variables must be indicated as precisely as possible, that is,including any relevant specifications, modifications or conditions. So instead of (3-a), (3-b) is preferred.

(3) a. *a difference in [*carbon concentration*] between the ocean surface and the deep waters
b. a difference in [*carbon concentration between the ocean surface and the deep waters*]

The choice is motivated by the assumption that, given a syntactic parse, it is usually easier to generalize a complex argument by stripping modifiers than the other way around. Variables are tagged with the label VARIABLE. We intend to distinguish different subclasses of variables, resulting in a more fine-graned categorisation of entities, in the near future. For now, we focus on annotation of the events and basic logic structure.

2.2.2. Change, Increase and Decrease

A change is an event in which the value of a quantitative variable is changing. The direction of change can be positive (increasing), negative (decreasing) or unspecified (either increasing or decreasing), but there must always be a clear cue in the text that the variable is involved in a change. This is referred to as the *trigger* for the event.

Examples of triggers for event types of change, increase and decrease are:

(4) a. [*regional changes in*] phytoplankton
b. [*addition of*] labile dissolved organic carbon
c. [*to slow down*] calcification in corals

Changes must apply to a variable; hence the text span in (5) does not trigger a change event.

(5) *marine primary production is sensitive to climate [*variability and change*]

Events of increase, decrease and undirected changes are tagged as INCREASE, DECREASE and CHANGE respectively. Events are related to variables through thematic roles, which specify the different participants in the event. Change events must always have a THEME role that is filled by the variable that is changing. Typical annotation examples are therefore:[3]

(6) a. [DECREASE reduced] [THEME calcite production]
b. [CHANGE significant changes in] [THEME surface ocean pH]

Change events can also function as Cause/Correlate events, as will be described in the next Section, in which case they take an AGENT or CO-THEME role as well.

2.2.3. Cause

Cause events involve a pair of changes where the first change causes the second change. Since a change event involves a changing variable, as its theme, causal events thus express a causal relation between two changing variables. The trigger of a cause event is annotated with a CAUSE tag. Triggers are often verbs, but can also be adjectives (*stimulatory*), adverbs (*therefore*) or subjunctive phrases (*due to*, *in response to*) or other phrasal expressions (*has an effect on*).
Cause events must always have two thematic roles: an AGENT identifying the cause and a THEME identifying the effect. Examples of cause events are:

(7) a. [AGENT rise in atmospheric CO2 levels] [CAUSE *causes*] [THEME significant changes in surface ocean pH]
b. [AGENT Fe(III) addition in the presence of GA (FeGA)] [CAUSE *gave*] [THEME higher Fe(II) concentration]
c. [AGENT diminished calcification] [CAUSE *led to*] [THEME a reduction in the ratio of calcite precipitation to organic matter production]

In many cases, a cause event and a change event share one and the same trigger, as in the following examples:

(8) a. [AGENT changes in the magnitude of total and export production] [CHANGE *can strongly influence*] [THEME atmospheric CO2 levels]
b. [THEME calcification and net primary production] [INCREASE *are significantly increased by*]

[2]The total amount of energy produced by marine organisms such as photosynthetic plankton.

[3]We use labeled brackets to denote entities, events or thematic roles, depending on the context of discussion.

[AGENT high CO2 partial pressures]

c. [AGENT addition of labile dissolved organic carbon] [DECREASE *reduced*] [THEME phytoplankton biomass]

In (8-a), *can strongly influence* serves as the cue for a change event with variable *atmospheric CO2 levels* as its theme. At the same time, it is the trigger for a cause event with agent *changes in the magnitude of total and export production* and theme *atmospheric CO2 levels*. In principle, both events can be annotated separately. However, in order to avoid a needlessly complex annotation, we chose to not annotate the cause event explicitly. Instead *changes in the magnitude of total and export production* is given the role of agent in the change event. Presence of an agent role suffices to infer the cause event. In other words, where there is both a cause event and a change event, we annotate the change event because it can be inferred from the presence of the agent role that a cause event is also being annotated. Two more instances of this pattern are shown in (8-b)[4] and (8-c).

2.2.4. Correlate

Correlate events involve a pair of changes where the first change correlates with the second change. Since a change event involves a changing variable, as its theme, correlate events thus express a correlation between two changing variables. That is, if one of them changes, the other changes along. Correlations have two roles, THEME and CO-THEME, both of which should be fulfilled by a change event (i.e. INCREASE, DECREASE or CHANGE). Examples of correlate events:

(9) a. [THEME reduced calcite production] [CORRELATE *was accompanied by*] [CO-THEME an increased proportion of malformed coccoliths]

b. [THEME carbon:nutrient ratio turns out to decrease] [CORRELATE *with*] [CO-THEME increasing mixed-layer depth and temperature]

c. Here we report [THEMe reduced calcite production] [CORRELATE *at*] [CO-THEME increased CO2 concentrations]

d. [CORRELATE *When*] [CO-THEME bacterial growth rate was limited by mineral nutrients], [THEME extra organic carbon accumulated in the system]

Notice that correlation can be triggered by a verb (9-a), a preposition (9-b-c) or an adverb/conjunction (9-d).
Statistically speaking, correlation is not a directional relation, in contrast to causation. That is, if a change in variable A is correlated with a change in variable B, then it follows that a change in variable B is correlated with a change in variable A. However, in discourse there is often a distinction between a variable of interest (the dependent variable) and a related variable (the independent variable). Thus even though strictly speaking there is no causal relation between the two variables, the text usually takes a particular perspective, suggesting one is more central than the other. By convention, the central variable is tagged as THEME, whereas the other one is tagged as CO-THEME. The rule of thumb is that the co-theme is syntactically the argument of a preposition (e.g. *with, at, under*) or an adverb/conjunction (e.g. *when*).

Occasionally correlations can hold between a change event and a variable, or even between two variables, rather than between two change events. In these exceptional cases, we assume the variable is interpreted as changing (i.e. as being part of an implicit change event), because it is involved in a correlate event. Two examples of this exceptional pattern are:

(10) a. [THEME:INCREASE Concentrations of DFe increased slightly] [CORRELATE *with*] [CO-THEME:VARIABLE depth in the water column]

b. [THEME:VARIABLE growth rates in the high-CO2-grown cells] [CORRELATE *were related to*] [CO-THEME:VARIABLE light level]

In (10-a), the role of co-theme is not taken by a change event, but by the variable *depth in the water column*. It is thus assumed that the depth in the water column is a changing variable in the correlation described. Similarly, (10-b) has both roles of the correlate event taken up by variables, which are therefore interpreted as subject to change.

2.2.5. Feedback

Feedback loops are an important concept in climate science. An example is that of the relation between rising temperature and methane release: a rise in temperature causes more permafrost to melt, which causes more release of methane in the atmosphere (a "green house" gas), which causes further rising of the temperature, and so on. However, explicit mentioning of feedback events in the text appears to be rare compared with the frequent occurrence of change events, so our proposal for annotation of feedbacks is currently based on only a couple of instances. Feedback events hold between two variables, filling the roles of THEME and CO-THEME, as exemplified below:

(11) our model suggests the existence of [+FEEDBACK *a positive feedback between*] [THEME temperature] and [CO-THEME atmospheric CO2 content]

Analogously to change events, feedback events can be positive (self-sustaining, self-enhancing), negative (self-stabilising, self-diminishing) or of unspecified polarity. Positive or negative feedback are annotated with an attribute whenever a trigger is present.

2.2.6. Referring expressions

Referring expressions such as anaphoric expressions (e.g. *it*, *this*) and underspecified definite descriptions (e.g. *the process*) are annotated only in so far as they play a thematic role in an event of interest. Consider the following narrative:

[4] The agent in example (8-b), i.e. *high CO2 partial pressures*, is arguably not an event but a state. However, we took the liberty to interpret this as *increasing CO2 partial pressures* in this context, which is in accordance with the interpretation of our domain expert.

(12) s1: Future shoaling of upper-mixed-layer depths will expose phytoplankton to [INCREASE increased] [THEME mean light intensities].
s2: [REFEXP/AGENT *This*] [CAUSE may cause] [DECREASE a widespread decline in] [THEME marine primary production]

A graphical representation of a slightly extended version of this example is shown in Figure 1. The first sentence contains an INCREASE event, which is referred to in the second sentence by means of the referring expression *This*, establishing it as the cause for the DECREASE event. Such referring expressions must therefore be resolved in order to deduce the rule that an increase in mean light intensities causes a decrease in marine primary production. In order to achieve this, they are tagged as REFEXP and connected with their antecedent by means of a COREF relation.

2.2.7. Combinations

Variables or events can be combined through conjunction or disjunction. Such combinations are labeled as AND or OR, where their constituents fill the role of PART. In (13-a), for example, the combination AND serves as the theme of the INCREASE event. Likewise, two increasing events are combined to serve as the theme in a causal event.

(13) a. [INCREASE increasing] [PART:VARIABLE mixed-layer depth] [THEME:AND and] [PART:VARIABLE temperature]
b. [CAUSE gave] [PART:INCREASE higher] [THEME Fe(II) concentration] [THEME:AND and] [PART:INCREASE higher] [THEME growth rate of phytoplankton]

The alternative option in (13-a) is to tag the whole combined phrase as a single variable. We chose not do so because coordination is a notoriously hard problem for syntactic parsers and any help from the annotation in resolving ambiguity should be exploited. Notice also that a similar option is not available in (13-b), as considering the whole combination as a single change event would result in loss of substantial information.

There are certain cases, like where an adjectival modifier modifies a conjunction of two variables, that can not be accommodated by the proposed annotation scheme. This is not a shortcoming of the Brat annotation tool, but a matter of trade-off between expressivity and complexity: covering these instances requires additional relations or events, which would further complicate the annotation process. However, judging from the sample texts annotated so far, these cases are rare.

2.2.8. Negation

Events can carry a negation attribute to account for examples such as:

(14) a. TaLFe [CORRELATE+NEG did *not* show any consistent trend with] depth
b. [CHANGE+NEG *No* differences] in cellular organic carbon:nitrogen ratios were observed

Triggers for negation are currently not explicitly annotated.

3. Rule extraction

The proposed annotation allows for automatic extraction of rules about the relations between quantitative variables. There are three main types of rules: causal rules, correlation rules and feedback rules.

Causal rules are of the type "If variable X changes, than variable Y changes". An example of such a rule and its source text is shown in Figure 1. The notation uses single arrows to denote changing variables, where '$\uparrow$' stands for 'increasing', '$\downarrow$' for 'decreasing' and '$\updownarrow$' for 'changing.' Parts of a combination are joined by '$\wedge$' or '$\vee$' and delimited by square brackets. A causal relation is denoted by the double arrow '$\Longrightarrow$'. Causal rules are basically extracted by looking for CAUSE events, taking their AGENT and THEME roles for cause and effect respectively. Notice that in Figure 1, interpreting combinations and resolving referring expressions to their antecedent takes some additional processing. Another source for causal rules is change events with both AGENT and THEME roles.

Correlation rules are of the type "Changes in variable X correlate with changes in variable Y", as exemplified in Figure 2. The curly arrow '$\rightsquigarrow$' is used to indicate the relation between a independent and a dependent variable. These rules are extracted from CORRELATE events, using their CO-THEME role as the independent variable (LHS of the rule) and their THEME role as the dependent variable (RHS of the rule).

Feedback rules, an example of which is shown in Figure 3, are of the form: "Changes in variable X feed back through changes in variable Y". The feedback relation is denoted by a double sided arrow '$\Longleftrightarrow$', optionally with a superscripted '+' or '-' for positive and negative feedback respectively.

Notice that conceptually a feedback relation is assumed to hold between changing events. However, often there is no explicit trigger for a change event present in the text. For example, in the annotation in Figure 3, both roles are filled by variables instead of change events. Such variables are therefore 'promoted' to change events during rule extraction, resulting in '$\updownarrow$ temperature' and '$\updownarrow$ marine primary production'. Similar promotions apply occasionally to variables in events of change, cause or correlation (cf. Section 2.2.4.).

4. Discussion

The annotation scheme proposed below seems a good candidate for the purpose of rule extraction. However, it has only been tried on a small set of abstracts and it remains to be seen how it holds up when applied to more text. To provide some indicative statistics, the pilot-corpus contains the following number of labels: 107 VARIABLE, 33 CHANGE, 82 INCREASE, 50 DECREASE, 20 CAUSE, 26 CORRELATE, 32 AND, 2 OR, 5 REFEXP and 2 NEGATION. Annotation of more text is required to settle certain corner cases and is likely to reveal additional issues. For example, the current scheme can not capture the fact that *ocean acidification* is an event, i.e., a decrease of the pH of the ocean water. If similar examples turn out to occur frequently, this may cause a revision of the annotation scheme. Inter-annotator agreement has not been measured so far. In addition to this,

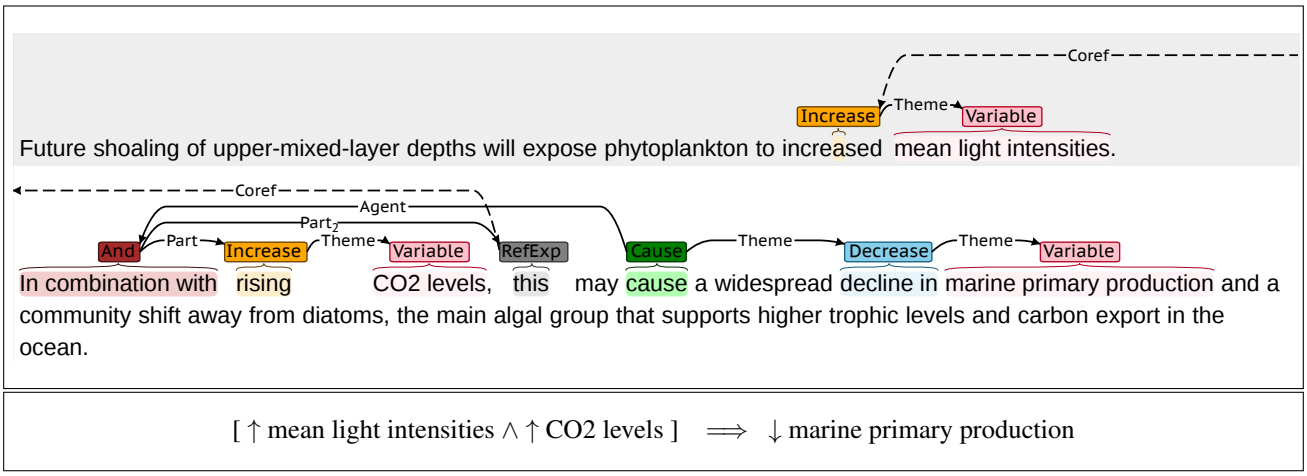


Figure 1: Example of a causal rule extracted from a pair of annotated sentences

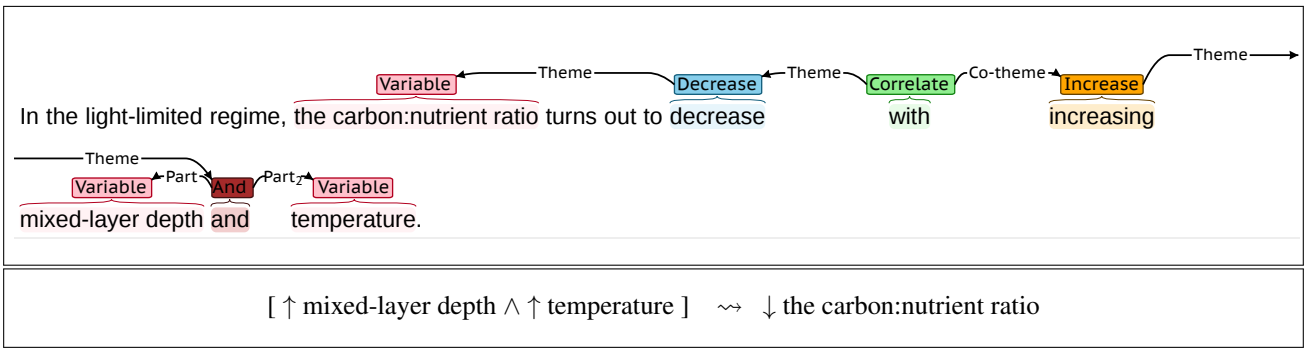


Figure 2: Example of a correlation rule extracted from an annotated sentence

we are also considering a type of evaluation in which extracted rules and their corresponding source texts are shown to domain experts, who are then asked to judge if the rule is entailed by the text.

Manual annotation is costly. There are at least two strategies which may reduce annotation time and costs. The first one is to bootstrap from existing extraction systems. Recent advances in *open information extraction*, where there is no predefined set of entities and relations, have resulted in open source systems like ReVerb (Fader et al., 2011) and it successor OpenIE. Banko and Etzioni (2008) claim that when the number of target relations is small, and their names are known in advance, an open IE system is able to match the precision of a traditional supervised extraction system, though at substantially lower recall. This suggests that at least a part of the annotation can be accelerated with the help of such tools.

A second strategy to reduce annotation costs involves the use of *active learning*, which is a training method for supervised learners that tries to obtain maximal performance gain with minimal annotation effort (Olsson, 2009). It is an iterative procedure, starting with a small amount of labeled data and a large amount of unlabelled data. In each iteration, a classifier is trained on the labeled data and subsequently applied to the unlabelled data. Only the most informative instances – e.g., those for which classification confidence is lowest – are passed on to a human annotator for manual annotation. These manually labeled instances are added to the training data and the procedure is repeated. Good results have been reported with the use of active learning, e.g. by (Gambäck et al., 2011).

The extracted rules expressing relations of correlation, causality or feedback between quantitative variables are intended to be used in knowledge discovery support systems. One use case is to search for other variables directly related to a certain variable of interest. For example, find all processes that affect or are affected by a rise in atmospheric CO2 level. The variable in question may be expressed in many different ways though, for example, as *CO2*, *atmospheric CO2*, *CO2 concentrations* or *CO2 partial pressures*, but not as *CO2 levels in oceanic surface waters* or *the distribution of CO2 between the atmosphere and the ocean*. Simple string matching between the variables in queries to those in rules will given limited recall and precision. Related to this is the issue of differences in terminology across research fields. For instance, *export production* and *biological pump* are different terms, used by chemists and biologists respectively, for the same process of carbon cycling in the oceans. One possible strategy to cope with this issue is to have a more fine-grained categorisation of entities, allowing different surface realisations to be mapped to the same underlying domain concept. This would allow more general rules to be extracted, and could also be beneficial in helping to bootstrap lexical resources.

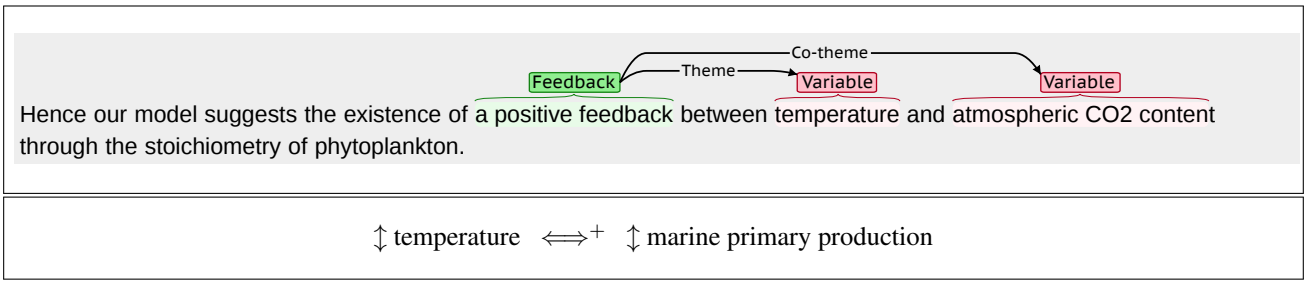


Figure 3: Example of a feedback rule extracted from an annotated sentence

Ultimately all relevant entities may be normalised by linking them to a unique concept in a domain ontology (Bada et al., 2012). However, whereas the concepts of interest in biomedicine are relatively well understood – including such entities as cells, proteins and genes – and covered by widely used ontologies, such common ground currently seems to lack in climate, marine and environmental science science.

A different but related problem is exemplified in correlation rule (15-b) extracted from the second part of sentence (15-a).

(15) a. Concentrations of DFe increased slightly with depth in the water column, while that of TaLFe did not show any consistent trend with depth.

b. ¬ [↕ depth ⇝ ↕ that of TaLFe]

The problem is that *depth* in (15-b) is too general and should in fact be linked to *depth in the water column* for proper interpretation. Likewise, *that of TaLFe* should be interpreted as *concentrations of TalFE*. This illustrates the need for coreference resolution and more general, linking of subsequent mentions of the same entity in the text, a notoriously hard task in NLP.

Apart from search, another use case for extracted rules is to generate potential hypotheses about indirect relations between variables or feedback loops among them. This can be accomplished by chaining together two or more rules, matching the change event on the right-hand-side of one rule to a similar change event on the left-hand-side of another rule. Matching gives rise to the same problems discussed above, i.e., different ways of referring to the same entity. In addition, there is the issue of context-dependency. Most rules are not universally applicable, but only apply under certain conditions in a particular context. For example, a rule may be limited in scope to certain biological species or organisms, a particular geographical region or historical time period, subject to a given assumption (*only if* ...), etc. This is related to initiatives for annotating meta-knowledge such as confidence level (fact vs. conjecture), source (resulting from observation vs. analysis) or origin (present or cited work) as in (Thompson et al., 2011). Proper modelling of rule context would require a rather deep understanding of the whole text. Although we acknowledge the importance of conditions on events, we intend to leave their annotation to a later stage. For now, we plan to leave this to the user by offering facilities in the user interface to quickly inspect the source text for each rule.

Inference with rules may be further enhanced by exploiting domain knowledge. For example, given an ontology which contains the fact that *diatoms* are a kind of *phytoplankton*, rules containing either of the terms may be generalised by substituting the hypernym or specialised by substituting the hyponym. In a similar vein, rules can be generalised by removing specifiers, modifiers or parts of a conjunction. Whether or not this constitutes valid inference seems connected to recent developments in textual entailment, in particular work on natural logic (MacCartney and Manning, 2008).

5. Conclusion

An annotation scheme was proposed to capture events of change, cause, correlation and feedback, as well as the entities involved in them, in the cross-disciplinary fields of climate science, marine science and environmental science. It was shown that rules about the relation between changing processes can be automatically extracted from annotated text. Follow-up work will involve annotating more text, as well as measuring inter-annotator agreement and rule adequacy. Simultaneously, tools for automatic annotation will be developed. Future work will also address normalisation of entities, tracking of entity mentions, modelling of rule context and combination with domain knowledge.

6. Acknowledgements

Financial aid from the European Commission (OCEAN-CERTAIN, FP7-ENV-2013-6.1-1; no: 603773) is gratefully acknowledged. We thank the reviewers for their valuable comments.

7. References

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13(1):161+, July.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36, Columbus, Ohio, June. Association for Computational Linguistics.

Oren Etzioni. 2011. Search needs a shake-up. *Nature*, 476(7358):25–26, August.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Björn Gambäck, Fredrik Olsson, and Oscar Täckström. 2011. Active learning for dialogue act classification. In *Twelfth Annual Conference of the International Speech Communication Association*.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong H. Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 619–630, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dimitar Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. 2008. Literature-Based Knowledge Discovery using Natural Language Processing. In Peter Bruza and Marc Weeber, editors, *Literature-based Discovery*, volume 15 of *Information Science and Knowledge Management*, chapter 9, pages 133–152. Springer, Heidelberg, Germany.

J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpusa semantically annotated corpus for biotextmining. *Bioinformatics*, 19(suppl 1):i180–i182, July.

Jin D. Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Robert K. Lindsay and Michael D. Gordon. 1999. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, pages 574–587.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 521–528, Stroudsburg, PA, USA. Association for Computational Linguistics.

Claudiu Mihăilă and Sophia Ananiadou. 2013. Recognising discourse causality triggers in the biomedical domain. *Journal of bioinformatics and computational biology*, 11(06).

Claudiu Mihaila, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. BioCause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14(1):2+.

F. Olsson. 2009. A literature survey of active machine learning in the context of natural language processing. Technical Report Tech. Rep. T2009, SICS, Stockholm, Sweden.

Chaveevan Pechsiri and Rapepun Piriyakul. 2010. Explanation knowledge graph construction through causality extraction from texts. *Journal of Computer Science and Technology*, 25(5):1055–1070.

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):188+, May.

Thomas C. Rindflesh and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. *J. of Biomedical Informatics*, 36(6):462–477, December.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, April. Association for Computational Linguistics.

Thomas F Stocker, Q Dahe, and Gian-Kasper Plattner. 2013. Climate change 2013: The physical science basis. *Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Summary for Policymakers (IPCC, 2013)*.

Don R. Swanson. 1986a. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1):7–18.

Don R. Swanson. 1986b. Undiscovered public knowledge. *The Library Quarterly*, 56(2):pp. 103–118.

Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2011. Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12(1):393+, October.

Marc Weeber, Henny Klein, Lolkje T. de Jong van den Berg, and Rein Vos. 2001. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557.

Sine Zambach and Tine Lassen. 2010. A lexical framework for semantic annotation of positive and negative regulation relations in biomedical pathways. In Nigel Collier, Udo Hahn, Dietrich R. Schuhmann, Fabio Rinaldi, Sampo Pyysalo, Nigel Collier, Udo Hahn, Dietrich R. Schuhmann, Fabio Rinaldi, and Sampo Pyysalo, editors, *Semantic Mining in Biomedicine*, volume 714 of *CEUR Workshop Proceedings*. CEUR-WS.org.

The *Quaero* French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization

Aurélie Névéol[1] **Cyril Grouin**[1] **Jeremy Leixa**[2]
Sophie Rosset[1] **Pierre Zweigenbaum**[1]
[1]CNRS, UPR 3251, LIMSI [2]ELDA
91403 Orsay, France 75013 Paris, France
{firstname.lastname}@limsi.fr leixa@elda.org

Abstract

A vast amount of information in the biomedical domain is available as natural language free text. An increasing number of documents in the field are written in languages other than English. Therefore, it is essential to develop resources, methods and tools that address Natural Language Processing in the variety of languages used by the biomedical community. In this paper, we report on the development of an extensive corpus of biomedical documents in French annotated at the entity and concept level. Three text genres are covered, comprising a total of 103,056 words. Ten entity categories corresponding to UMLS Semantic Groups were annotated, using automatic pre-annotations validated by trained human annotators. The pre-annotation method was found helful for entities and achieved above 0.83 precision for all text genres. Overall, a total of 26,409 entity annotations were mapped to 5,797 unique UMLS concepts.

Keywords: Corpus annotation; Named Entity Recognition; Entity Normalization

1. Introduction

A vast amount of information in the biomedical domain is available as natural language free text. Over the past decades, substantial research efforts have addressed the development of methods and tools to automatically process biomedical free-text and provide organized and structured representations of domain knowledge contained in the literature and in clinical documents. Much of this work relied on manually annotated corpora of biomedical texts written in English, such as GENETAG (Tanabe et al., 2005), GENIA (Kim et al., 2003) or corpora available through shared tasks such as i2b2 (Savova et al., 2011), BioCreative (Arighi et al., 2011) or BioNLP (Bossy et al., 2013). However, an increasing number of documents in the field are written in languages other than English. Therefore, it is essential to develop resources, methods and tools that address Natural Language Processing in the variety of languages used by the biomedical community.

While it is generally agreed that domain knowledge is expressed in text through *mentions* or *named entities* that may refer to domain concepts captured in domain terminologies or ontologies, the numerous annotated resources available for the biomedical domain show that there is no consensus on the representation and annotation of entities: what are the entities of interest? Should they be defined through meaning, syntax, both? What level of granularity should be taken into account? In this work, we chose to annotate a range of high-level semantic categories, in line with recent work addressing named entity recognition (Rebholz-Schuhmann et al., 2013) and corpus development (Albright et al., 2013) in the biomedical domain.

We produced corpora and annotation guidelines for named entities which are both hierarchical and compositional[1] (Grouin et al., 2011), and which we used in contrastive studies of news texts in French (Rosset et al., 2012). In this definition, in addition to the hierarchy and compositionality, the inclusion of entities covers antonomasia and metonymy but also takes into account cases of entities that are built upon other entities. This inclusion process seemed particularly interesting. Specifically, in this definition, complex entities can be decomposed into several simpler entities that may be of different categories.

2. Material and Methods

2.1. Corpus selection

A selection of text comprising relevant biomedical entities was made using three different types of documents: information on marketed drugs from the European Medicines Agency (EMEA),[2] titles of research articles indexed in the MEDLINE database,[3] and patents registered with the European Patent Office (EPO).[4] All three sources were recently used in an international challenge for cross-lingual medical named entity recognition, CLEF-ER (Rebholz-Schuhmann et al., 2013). We decided to use all three corpora because they cover three different genres of biomedical documents, and the availability of the documents in at least one language other than French (English for MEDLINE, English and German for EPO, several European languages for EMEA) provides an opportunity for a richer corpus.

Our goal in selecting documents was to obtain a sample with a few thousand annotations per semantic category in order to have a good representation of how these types of concepts were referred to in biomedical text. Based on manual annotation of a small sample of each text type, we randomly selected 2,500 MEDLINE titles, 13 EMEA documents and 25 EPO patents. To ensure the relevance of the selected patents to the biomedical domain, patents were selected among those containing at least one of the words "*maladie*" (disease) and "*médicament*" (drug).

[1]Corpora, guidelines and tools are available through ELRA under references ELRA-S0349 and ELRA-W0073.

[2]http://opus.lingfil.uu.se/EMEA.php

[3]http://www.ncbi.nlm.nih.gov/pubmed/

[4]http://www.epo.org/

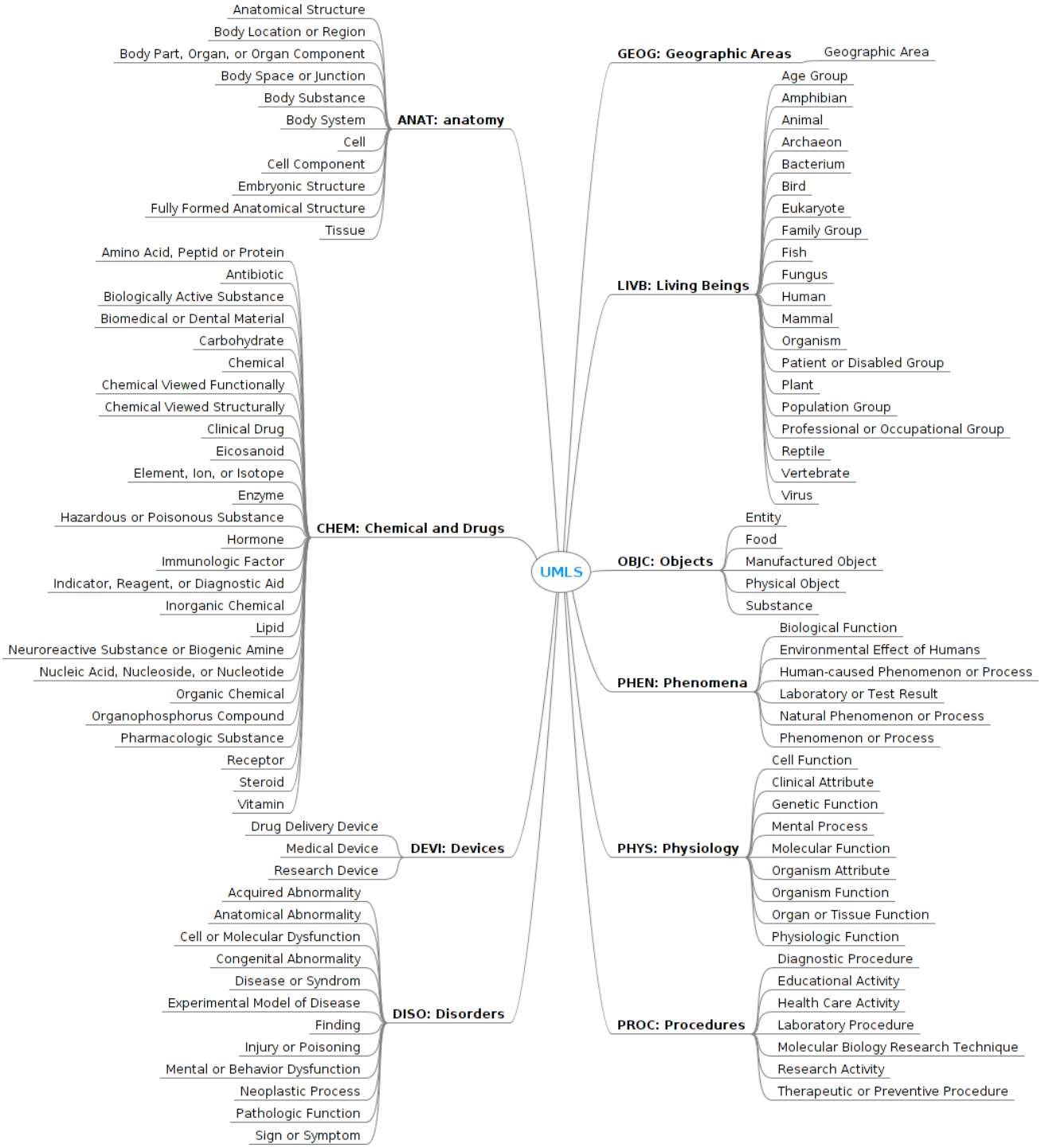


Figure 1: Annotated Entities

2.2. Annotation Guidelines

2.2.1. Entities are defined using the UMLS

The entities are defined based on the UMLS® (Unified Medical Language System®). The Metathesaurus unifies concepts from several dozen terminologies in the biomedical domain, including linked terms and relations. The Semantic Network comprises Semantic Types and Semantic Relations, which are organized hierarchically. The 134 Semantic Types can be clustered into 15 Semantic Groups (McCray et al., 2001). Each concept in the UMLS is assigned a Concept Unique Identifier (CUI), a set of terms and one or more Semantic Types. Semantic Groups were designed so that each concept could be assigned to only one semantic category (Bodenreider and McCray, 2003). Figure 1 shows the subset of entities defined according to the UMLS Semantic Groups and Semantic Types used in this work. The annotation task was carried out at two granularity levels: Semantic Groups and Concepts. The annotators used freely available tools for accessing the UMLS in

French[5] and in English[6] in order to assess whether a mention in text referred to a specific Semantic Group and to identify the specific concept.

2.2.2. Annotations are comprehensive

The goal of the annotation task was to provide a resource that would be as complete and comprehensive as possible. The guidelines applied the principles of *Quaero* annotations in the general domain, which imply that a complex entity can be decomposed into simpler entities that may belong to different categories. The main guidelines were as follows:

- If a mention can refer to more than one Semantic Group, all the relevant Semantic Groups should be annotated. For instance, the mention "*récidive*" (recurrence) in the phrase "*prévention des récidives*" (recurrence prevention) should be annotated with the category "DISORDER" (CUI C2825055) and the category "PHENOMENON" (CUI C0034897);
- If a mention can refer to more than one UMLS concept within the same Semantic Group, all the relevant concepts should be annotated. For instance, the mention "*maniaques*" (obsessive) in the phrase "*patients maniaques*" (obsessive patients) should be annotated with CUIs C0564408 and C0338831 (category "DISORDER");
- Entities which span overlaps with that of another entity should still be annotated. For instance, in the phrase "*infarctus du myocarde*" (myocardial infarction), the mention "*myocarde*" (myocardium) should be annotated with category "ANATOMY" (CUI C0027061) and the mention "*infarctus du myocarde*" should be annotated with category "DISORDER" (CUI C0027051);
- Discontinuous entities should be annotated separately. For instance, in the phrase "*maladies rares et chroniques*" (rare and chronic diseases) the entity "*maladies rares*" (rare diseases) should be annotated with the category "DISORDER" (CUI C0678236) and the entity "*maladies chroniques*" (chronic diseases) should be annotated with the category "DISORDER" (CUI C0008679).

2.3. Annotation Development

This section describes the annotation process performed by the human annotators. Two people were involved in the task over the course of 4 weeks: the project manager (JL) and one annotator specifically recruited for the project. Both annotators are native French speakers with a good command of English. When working out the workload distribution, we found that an expert annotator and a beginner could reasonably manage to annotate the medical corpus.
Before working on the medical corpus, the beginner annotator was trained for the annotation process on files taken from another (general) corpus. As part of the training, this annotator started working on the general corpus to practice applying the general annotation principles. Then, switching to the pre-annotated medical corpus was seamless because the annotation guidelines were very similar. The main difference was the use of the UMLS and PTS websites, and their role in the annotation process:

- For each Named Entity in the text (pre-annotated or not), search the corresponding concept in the UMLS database;
- If a suitable concept is found in the UMLS, check whether its category is listed in the annotation manual;
- If the Named Entity's category is of interest, report the associated CUI;
- Create a complete annotation for the Entity.

For the annotation task, the three sub-corpora (EMEA, MEDLINE and EPO) were divided equally between the annotators. In order to get used to the annotation process, the human annotators decided to begin with a few files from the MEDLINE sub-corpus (10 per annotator, out of 2,500), before working on the whole EMEA sub-corpus.
The tools used for the annotation task were Xemacs and dedicated configuration files enabling the processing of embedded annotations files from the medical corpus. As mentioned earlier, the medical corpus was supplied to the human annotators with a set of pre-annotations automatically obtained based on the method described in Bodnari et al. (2013). Table 1 shows a sample file with the original text content, the pre-annotated content supplied to the human annotators with embedded annotations and the annotations they finally produced. Note that due to time constraints, the documents were supplied to the human annotators without prior tokenization. This had an impact on the annotation task: because the Xemacs tool is only able to create annotations at the token level, in many cases, punctuations were included in the annotations done by the annotators, as seen in the "*cancer intestinaux.*" (intestinal cancer) example in Table 1, where the final full stop is part of the annotation.
A small portion (about 5%) of the corpus was annotated independently by both annotators, in order to calculate Inter Annotator Agreement (IAA). The first files from the MEDLINE sub-corpus were annotated in collaboration between the annotators, in order to discuss any annotation issues early on. The remaining files were annotated independently, but annotators still met once a week to discuss their issues and share comments on the annotation experience.

2.4. Corpus quality assurance and formatting

In order to ensure high-quality annotations, several quality assurance steps were included in the annotation process.
About 5% of the corpus was selected from the EPO and EMEA sub-corpora to be annotated by both the human annotators, and Inter Annotator Agreement (IAA) was computed (in terms of Kappa and F-measure) at the entity level for this sample. Additionally, two samples of 100 MEDLINE titles were selected randomly and an annotator from

[5]The Portail Terminologique de Santé (PTS) developed by the Rouen city hospital (`http://pts.chu-rouen.fr`)

[6]The National Library of Medicine Metathesaurus Browser (`https://uts.nlm.nih.gov/metathesaurus.html`)

Plain document	Facteurs de croissance et cancers intestinaux.
English translation	Growth factors and intestinal cancers.
Pre-annotated document	Facteurs de croissance et <DISO CUI="C0346627"> cancers intestinaux. </DISO>
Annotated document	<CHEM CUI="C0018284"> Facteurs de <PHYS CUI="C18270"> croissance </PHYS> </CHEM> et <DISO CUI="C0346627"> <DISO CUI="C0027651"> cancers </DISO> <ANAT CUI="C0021853"> intestinaux. </ANAT> </DISO>

Table 1: A sample MEDLINE document (PMID 1421706) with the automatic pre-annotations supplied to the human annotators, and the annotations they produced

the curator team who contributed to writing the guidelines (AN) revised the annotations supplied by the human annotators for these samples. IAA was computed between the human annotations version and curator-revised version of the samples.

A list including the most frequently encountered CUIs (such as most common body parts and diseases, age groups) was compiled so that both annotators would tag the most common Named Entities the same way.

At the end of the annotation process, the most common entities (such as "*syndrome*" (syndrome) and "*maladie*" (disease)) were pulled and the consensus category and CUI for each entity was enforced using Search and Replace method, when no ambiguity was possible. The consistency of CUI-category assignment for each entity was also checked automatically based on UMLS data: cases where, according to the UMLS, the CUI assigned to an entity did not belong to a Semantic Type corresponding to the assigned category (Semantic Group) were automatically flagged for correction. For example, when processing the entity "*<GEOG CUI="C0331677"> Paris </GEOG>*", the annotation of "*Paris*" with CUI C0331677 was flagged because CUI C0331677 (Genus Paris) is associated with the semantic group LIVB in the UMLS. This annotation could then be revised to the correct CUI C0030561 because the text referred to the city of Paris. The excerpt shown in Table 1, shows a case where the CUI C18270 was flagged because of a typo error. This annotation could then be revised to the correct CUI C0018270.

To improve the quality and accessibility of the annotations, the annotated corpus was post-processed for tokenization.

3. Results

3.1. Corpus statistics

Table 2 presents general corpus statistics including the number of tokens, the number of annotated entities, and the number of unique CUIs in the corpus.

Table 3 presents our evaluation of the performance of entity pre-annoation. We computed inter-annotator agreement scores between the pre-annotated (automatic) and final (revised by humans) versions of each corpus. We computed the Kappa coefficient (lower bound) and F-measure (higher bound), as defined in Grouin et al. (2011).

	EMEA	**MEDLINE**	**EPO**	**All**
Tokens	58,874	23,647	20,537	103,057
Pre-annotations				
Entities (all)	7,280	1,692	1,662	10,634
Entities (unique)	1,672	1,194	305	3,009
CUIs (all)	12,098	3,207	2,211	17,516
CUIs (unique)	1,653	1,879	378	3,325
Final corpus				
Entities (all)	12,761	8,781	4,865	26,407
Entities (unique)	2,839	5,600	960	8,460
CUIs (all)	12,647	8,767	4,867	26,281
CUIs (unique)	1,807	4,156	759	5,796

Table 2: Overview of the *Quaero* Medical corpus

Metric	EMEA	MEDLINE	EPO
Kappa	0.361	0.142	0.187
F-measure	0.595	0.303	0.428
Precision	0.831	0.937	0.841
Recall	0.463	0.181	0.287
Correct #	5,906	1,585	1,398
Insert #	6,261	7,123	3,394
Delete #	610	38	192
Substitution #	588	69	72

Table 3: Inter-annotator agreement scores and annotation comparisons between pre-annotated and final versions of each corpus

Table 4 presents the specific distribution of entity annotations over the ten Semantic Groups in each sub-corpus, in the pre-annotated version supplied to the annotators and the final version revised by annotators.

3.2. Inter-Annotator Agreement

Inter-annotator agreement was computed between the annotators and a domain expert (one of the guidelines writers) on three random samples of 100 MEDLINE titles. For the first sample, both the annotators and the expert worked from the pre-annotated documents. The analysis of annotation differences was used to consolidate the annotation guidelines. Due to time constraints, for the other two sam-

	EMEA				MEDLINE				EPO			
	Pre-annotated		*Final*		*Pre-annotated*		*Final*		*Pre-annotated*		*Final*	
	Total	%	Total	%	Total	%	Total	%	Total	%	Total	%
ANAT	461	6.33	1,031	8.08	150	8.87	1,464	16.67	197	11.85	323	6.64
CHEM	2,330	32.01	4,696	36.80	286	16.90	1,028	11.71	771	46.39	2,445	50.26
DEVI	84	1.15	340	2.66	3	0.18	126	1.43	16	0.96	256	5.26
DISO	2,608	35.82	2,698	21.14	838	49.53	2,825	32.17	171	10.29	381	7.83
GEOG	37	0.51	98	0.77	90	5.32	113	1.29	0	0.00	0	0.00
LIVB	991	13.61	1,370	10.74	283	16.73	899	10.24	183	11.01	315	6.47
OBJC	355	4.88	415	3.25	10	0.59	97	1.10	225	13.54	310	6.37
PHEN	21	0.29	142	1.11	3	0.18	160	1.82	30	1.81	248	5.10
PHYS	293	4.02	604	4.73	29	1.71	438	4.99	69	4.15	325	6.68
PROC	100	1.37	1,367	10.71	0	0.00	1,631	18.57	0	0.00	262	5.39
All	7,280	100.00	12,761	100.00	1,692	100.00	8,781	100.00	1,662	100.00	4,865	100.00

Table 4: Number of annotations for each category in both the pre-annotated and the final corpora

ples considered, the expert revised the annotators' final annotations.

IAA for entities. We consider agreement on entities at the mention and category level. This measure evaluates whether annotators selected the same text span and assigned the same category, among the ten Semantic Groups listed in the guidelines. Due to the difficulty of evaluating the number of markable entities according to Hripcsak and Rothschild (2005), Grouin et al. (2011), and Fort et al. (2012), IAA was assessed using F-measure computed on strict boundary and type match. For the first set, agreement on entities was 62% Kappa and 77% F-mesure. For the other two sets, the agreement was 92% and 90% F-measure.

IAA for CUIs. The IAA for CUIs was computed using F-measure at the sentence level. Each sentence was indexed using the set of CUIs assigned by the annotators to entities in the sentence. IAA was assessed using F-measure computed on strict CUI match. For the first set, agreement on CUIs was 66% F-mesure. For the other two sets, the agreement was 91% F-measure.

4. Discussion

4.1. Quality of the annotations

4.1.1. Influence of pre-annotation

The annotators felt that the pre-annotation was useful at the entity level. They felt that little revision was needed to the terms selected by the automatic system. Changes mostly consisted in adding entities missed by the system. However, the annotators did not like the CUI suggestions provided automatically. For many entities, the CUI suggestion consisted in a very long list and it was often felt that looking for a CUI independently from the suggestion was a faster method for doing the entity normalization.

This perception of the pre-annotation performance is reflected in Table 3, which shows good precision and poor recall for entity pre-annotation, and Table 2, which shows that the number of CUI per entity is much higher in the pre-annotations vs. final annotations.

Both Table 2 and 3 show significant differences between the pre-annotations and the final annotations, indicating that substantial work was required to prepare the corpus.

4.1.2. Annotation challenges

We believe that the annotation task addressed in this work was particularly difficult due to the combination of large entity coverage (ten types of entities), large target vocabulary for normalization (the entire UMLS), and the specialized nature of the texts in the corpus. We elaborate on more specific difficult points below.

Annotation time. Annotation time was about 0.85 annotation per minute for a domain expert revising the pre-annotated documents, as estimated based on one sample of 100 MEDLINE titles. Annotation time reached an estimated 2.2 annotations per minute for a domain expert revising annotators' work (based on two samples of 100 MEDLINE titles). This reflects on the difficulty of the task: annotation time was overall quite long, though it was faster to revise annotators' work, compared to the automatic pre-annotations. The manual annotations were of better quality; a smaller number of changes and CUI checks were needed. Annotation time was significantly longer than for broadcast news; in Rosset et al. (2013), we estimated annotation time ranged between 35 annotations per minute (for expert annotators) and 13 annotations per minute (for novice annotators). Note that our task covered both entity annotation and entity normalization, while our previous study (Rosset et al., 2013) only covered entity annotation.

Language barrier. While the PTS is a precious source of biomedical terminology in French, many concepts are only linked to terms in English. Because the annotation instructions stated that an entity should be annotated only if a corresponding concept is found in the UMLS, the annotators may have failed to find a relevant concept in the Metathesaurus browser because they did not know how to express the concept in English. In spite of these difficulties, one third of the unique CUIs assigned to the entities in the corpus (1,939 out of 5,796) do not have a French term associated with them, including the most frequent CUI in the corpus C0087111 "Therapeutic procedure" (*Traitement*).

Complex entities. The annotation of complex entities was a source of inter-annotator disagreement, as one annotator sometimes omitted to annotate either one component or the complex entity itself. For example, the

phrase "*syndrome de Moschcowitz*" (Moschowitz's syndrome) was annotated with category "DISORDER" (CUI C0034155) by two annotators, however only one annotator annotated the phrase "*syndrome*" (syndrome) with category "DISORDER" (CUI C0039082). In another case, in the phrase "*Anévrisme de l'aorte thoracique*" (thoracic aorta aneurysm), both annotators annotated "*Anévrisme*" (aneurysm) with the category "DISORDER" (CUI C0002940) and "*aorte thoracique*" (thoracic aorta) with the category "ANATOMY" (CUI C1281571), but only one of them annotated the whole phrase with the category "DISORDER" (CUI C0162872). In both cases, all the annotations were relevant.

Completeness. Annotating all the relevant entities was a challenge. One reason for this difficulty is the lack of prior knowledge of biomedical terminologies for some of the annotators. Another reason is that all concepts are not covered in the terminologies; for instance many Geographical Locations or Living Being concepts are not covered so annotators may have been tempted to assume that a concept did not exist. For example, "*Chicago*" should be annotated as "GEOGRAPHIC" entity (CUI C0008044). However, there is no concept in the UMLS for "*Detroit*", which should not be annotated.

4.2. Contributions of this work

This work reports on the on-going development of a substantial high-quality ressource to the community in a language other than English, viz. French. It provides a unique annotated corpus covering three types of biomedical text and ten Semantic Groups with discontinuous annotations and overlapping annotations. Interestingly, about one third of the concepts assigned to entities do not currently have a French term linked to them through the current versions of the terminologies. This shows the potential contribution of this corpus for terminology development. Furthermore, the corpus provides a semantic characterization of three types of biomedical text, which has adds to previous sub-domain studies (Mihğailğa et al., 2012). It can be seen from Tables 2 and 4 that the profiles of each sub-corpus are quite different in terms of semantic types represented, variety and redundancy of concepts represented.

4.3. Comparison to other work

The choices made in the design of this annotation work were guided by the goal of developing a comprehensive annotated corpus. In doing so, we built on the experience of previous annotation efforts both within and outside of the biomedical domain.

Overall annotation methodology. Entity annotation and concept mapping were performed together (Similar to CRAFT (Bada et al., 2012) but unlike SHARP (Savova et al., 2012) and NCBI disease corpus (Doğan et al., 2014) where entity annotation was a separate task, performed prior to normalization). The annotation guidelines explicitly stated that if an entity that belonged to the annotated categories was found but could not be linked to a UMLS concept, no annotation was to be created.

Normalization method. The entities annotated were to be normalized using concepts in the entire UMLS Metathesaurus (more than one million concepts). Unlike SHARP and NCBI corpora, we did not limit normalization to one or two vocabularies such as SNOMED, MeSH or OMIM (several thousand concepts). When no suitable concept was found in the UMLS to normalize a given entity, similarly to CRAFT guidelines, we chose to not create an entity annotation at all, instead of assigning a "CUI-less" concept (SHARP) or assigning a close concept or combination of concepts (NCBI). The Bacteria Biotope Corpus from BioNLP 2013 normalized habitat entities to the OntoBiotope ontology (several thousand concepts). When an entity corresponded to a concept that was not in the OntoBiotope ontology, it was linked to the closest general concept.

Semantic coverage. The *Quaero* medical corpus covers entities that belong to 10 UMLS Semantic Types. This makes it the largest annotated corpus in a language other than English for the biomedical domain. The MANTRA initiative also used the same three text types, but it currently offers silver standard (i.e., automatically obtained) annotations and only covers 9 Semantic Groups. Other corpus in English provide annotations for some of the Semantic Groups covered in the *Quaero* medical corpus. For instance, the SHARP/CLEF e-Health corpus and NCBI disease corpus cover the "DISORDER" group.

Nested entities. The i2b2 2011 (Uzuner et al., 2012; Savova et al., 2011) and BioNLP 2013 (Bossy et al., 2013) corpora feature nested entities. However, for i2b2, nesting was limited to two entity types (anatomy and medical problems); for BioNLP, there were only 3 entity global categories (bacteria, geographical, habitat). Entity nesting was permitted, including for same-type entities. In our corpus many nested entities covered the entire span of a larger entity while in the BioNLP corpus, a nested entity often covered only a small portion of the span of a larger entity.

5. Conclusion and perspectives

We presented the development of a large annotated corpus for biomedical texts in French. The annotation effort was guided by the desire to provide the community with a comprehensive entity resource. We relied on the UMLS for category definition and entity normalization.

Our experience shows that annotating a large corpus with a somewhat complex scheme is a hard task, both technically and cognitively. Recent reviews of annotation tools (Neves and Leser, 2012) and annotation frameworks (Comeau et al., 2013) are testimony to the technical issues. Annotation quality was assessed throughout the project by computing inter-annotator agreement on randomly selected corpus samples.

Further quality control is being performed on the final corpus in order to resolve inconsistencies and formatting issues. We are also planning to convert the corpus to the BioC format (Comeau et al., 2013) in order increase its accessibility and usability. The corpus will be used for organizing a challenge/shared task on entity annotation and normalization, and released to the community.

Acknowledgments

This work has been partially funded by OSEO under the Quaero program and by the French ANR Cabernet project (ANR-13-JS02-0009).

6. References

Albright, D., Lanfranchi, A., Fredriksen, A., 4th Styler, W., C, C. W., Hwang, J., Choi, J., Dligach, D., Nielsen, R., Martin, J., Ward, W., Palmer, M., and Savova, G. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, 20(5):922–930.

Arighi, C., Lu, Z., Krallinger, M., Cohen, K., Wilbur, W., Valencia, A., Hirschman, L., and Wu, C. (2011). Overview of the BioCreative III workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., and Hunter, L. E. (2012). Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(161).

Bodenreider, O. and McCray, A. T. (2003). Exploring semantic groups through visual approaches. *J Biomed Inform*, 36(6):414–32.

Bodnari, A., Névéol, A., Uzuner, O., Zweigenbaum, P., and Szolovits, P. (2013). Multilingual named-entity recognition from parallel corpora. In *Proc of the CLEF 2013 Evaluation Labs and Workshop*.

Bossy, R., Golik, W., Ratkovic, Z., Bessières, P., and Nédellec, C. (2013). BioNLP shared task 2013 – an overview of the bacteria biotope task. In *Proc of BioNLP Shared Task 2013*, pages 161–169, Sofia, Bulgaria. Association for Computational Linguistics.

Comeau, D. C., Doğan, R. I., Ciccarese, P., Cohen, K. B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., Valencia, A., Verspoor, K., Wiegers, T. C., Wu, C. H., and Wilbur, W. J. (2013). BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, page bat064.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *J Biomed Inform.*, 47:1–10.

Fort, K., François, C., Galibert, O., and Ghribi, M. (2012). Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proc of LREC*, Istanbul, Turkey. European Language Resources Association (ELRA).

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011). Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proc of Linguistic Annotation Workshop (LAW-V)*, pages 92–100, Portland, Oregon, USA. Association for Computational Linguistics.

Hripcsak, G. and Rothschild, A. S. (2005). Technical brief: Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc*, 12(3):296–298.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *BMC Bioinformatics*, 19(Suppl 1):i180–2.

McCray, A. T., Burgun, A., and Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. In *Proc of MedInfo*, volume 10, pages 216–20.

Mihğailğa, C., Batista-Navarro, R. T., and Ananiadou, S. (2012). Analysing entity type variation across biomedical subdomains. In *Proc of Building and Evaluating Resources for Biomedical Text Mining Workshop*.

Neves, M. and Leser, U. (2012). A survey on annotation tools for the biomedical literature. *Brief Bioinformatics*, pages 523–47.

Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E. M., Bui, C., Hellrich, J., Lewin, I., Milward, D., Poprat, M., Jimeno-Yepes, A., Hahn, U., and Kors, J. A. (2013). Multilingual semantic resources and parallel corpora in the biomedical domain: the clef-er challenge. In *Proc of the CLEF 2013 Evaluation Labs and Workshop*.

Rosset, S., Grouin, C., Fort, K., Galibert, O., Kahn, J., and Zweigenbaum, P. (2012). Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proc of Linguistic Annotation Workshop (LAW-VI)*, pages 40–48, Jeju, Republic of Korea. Association for Computational Linguistics.

Rosset, S., Grouin, C., Lavergne, T., Ben Jannet, M., Leixa, J., Galibert, O., and Zweigenbaum, P. (2013). Automatic named entity pre-annotation for out-of-domain human annotation. In *Proc of Linguistic Annotation Workshop and Interoperability in Discourse (LAW-VII & ID)*, pages 168–177, Sofia, Bulgaria. Association for Computational Linguistics.

Savova, G., Chapman, W., Zheng, J., and Crowley, R. (2011). Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, 18(4):459–65.

Savova, G., Styler, W., Albright, D., Palmer, M., Harris, D., Zaramba, G., Haug, P., Clark, C., Wu, S., and Ihrke, D. (2012). SHARP template annotations: Guidelines. Technical report, Mayo Clinic.

Tanabe, L., Xie, N., Thom, L., Matten, W., and Wilbur, W. J. (2005). Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., and South, B. R. (2012). Evaluating thestate of the art in coreference resolution for electronic medical records. *J Am Med Inform Assoc*, 19(5):786–91.

BioC and Simplified Use of the PMC Open Access Dataset for Biomedical Text Mining

Rezarta Islamaj Doğan, W. John Wilbur and Donald C. Comeau
National Center for Biotechnology Information, National Library of Medicine,
Bethesda, MD 20894, USA
E-mail: Rezarta.Islamaj@nih.gov, wilbur@ncbi.nlm.nih.gov, comeau@ncbi.nlm.nih.gov

Abstract

High quality easily-accessible resources are crucial for developing reliable applications in the health and biomedical domain. At the same time, interoperability, broad use, and reuse are vital considerations when developing useful systems. As a response, BioC has recently been put forward as a convenient XML format to share text documents and annotations, and as an accompanying input/output library to promote interoperability of data and tools. The BioC approach allows a large number of different textual annotations to be represented, and permits developers to more easily and efficiently share training data, supportive software modules and produced results. Here we give a brief overview of BioC resources. We also present the BioC-PMC dataset as a new resource, which contains all the articles available from the PubMed Central Open Access collection conveniently packaged in the BioC format. We show how this valuable resource can be easily used for text-mining tasks. Code and data are available for download at the BioC site: http://bioc.sourceforge.net.

Keywords: interoperability, PubMed Central, biomedical natural language processing, BioC, annotations

1. Introduction

The BioCreative[1] challenge evaluations since 2003 have reflected community-wide efforts for evaluating text mining and information extraction systems, and have sought to promote research on the topic. A goal of these meetings has consistently been to make available both suitable information extraction systems that handle life science literature and suitable "gold standard" data for training and testing these systems (Arighi et al., 2011; Hirschman et al., 2005; Krallinger et al., 2008; Wu et al., 2012). The BioCreative IV meeting, held in October 2013, specifically addressed the goal of interoperability—a major barrier for wide-scale adoption of the developed text mining tools. As a solution, BioC[2] (Comeau, et al., 2013) is a simple XML format, specified by DTD, to share text documents and annotations. BioC is also a suite of software tools to read and write the BioC format in multiple common computer languages. The BioC annotation approach allows many different annotations to be represented, including sentences, tokens, parts of speech, and named entities such as genes or diseases. The BioC repository offers several corpora with biomedical data annotations in BioC format. In order to increase the usefulness of BioC, it is important to make more data available in the BioC format.

Much of the data used for biomedical text mining comes from PubMed. PubMed currently contains more than 23 million citations for biomedical literature, of which, 2.9 million are currently deposited to PubMed Central[3] (PMC) for free full-text access. Furthermore, the Open Access subset of PMC, which contains more than 700 thousand full-text articles, is made available for further redistribution and reuse under Creative Commons licences. This resource is very important to the biomedical text mining community, because traditionally access to full-text has been limited by copyright restrictions, preventing large scale processing. Many biomedical text mining and NLP algorithms are trained on PubMed data and there is evidence that more precise and detailed information could be obtained through processing of full-text articles. Many authors have studied, or pointed out as a conclusion of their study, the differences between data in abstracts and full-text article bodies (Blaschke & Valencia, 2001; Cohen & Hersh, 2005; Corney et al., 2004; Divoli et al., 2010; Hearst et al., 2007; Hirschman et al., 2012; Shah et al., 2003; Yu & Agichtein, 2003).

In this work, we present the PMC Open Access subset in BioC format. Our goal is to make the text of the articles easily accessible, by providing a simpler alternative to the general-purpose PMC XML, which has been designed to accommodate multiple publishers and allow a rich online display. By allowing easier access to textual information in PMC Open Access articles, we hope that the resource will become simpler to use, and the wealth of information it contains will be utilized more efficiently.

In the following sections we begin with a brief overview of BioC, describe available tools and enumerate some of the corpora that have been converted to the BioC format. This is followed by a description of the BioC-PMC corpus. Finally we present the results of an example application of the BioC-PMC corpus. All abbreviation definitions are extracted from BioC-PMC text and results are analysed and discussed.

[1] http://www.biocreative.org/
[2] http://bioc.sourceforge.net/
[3] http://www.ncbi.nlm.nih.gov/pmc/

2. The BioC format

The BioC DTD defines the syntax in which a document can contain text, annotations and relations. All data in BioC is organized in collections. Collections have attributes that identify their source, date and `keyfile`. The building unit of a collection is a document. Documents are identified by their unique identifiers and are composed of passages which may be further separated into sentences. Annotations, and relations between them, can be defined at the sentence or passage level. Every element in a BioC file may be described by units of information defined in key-value pairs termed `infons`.

Keyfiles in BioC are plain text documents compiled by the authors of the collection that explain important semantic details of the documents in the collection and the meaning of the tags or tag sets used in this data collection. A keyfile must accompany each BioC formatted collection. Different corpora or annotation sets that share the same semantics may reuse an existing keyfile. The keyfiles of the currently available BioC collections are available from the BioC website. In time, we believe, the most useful keyfiles will develop a life of their own, thus providing emerging standards that are naturally adopted by the community.

The base format of BioC is XML because XML is well known, well documented, and well implemented. Standard XML tools can be used when convenient. But to ease and facilitate BioC use for BioNLP researchers, BioC also includes libraries for reading data into and writing data out of data structures in a number of common languages, including C++, Java and Python. The main goals are achieving interoperability, allowing tool developers to spend their time on the task at hand, and minimizing the time spent on learning a new data format.

3. BioC tools

Teams participating in the BioCreative IV challenge have also contributed a number of text and natural language processing tools which handle the BioC format as well as conversion tools that help convert corpora in other formats into BioC and vice-versa (Comeau et al., 2013; Islamaj Dogan et al., 2013; Jimeno Yepes et al., 2013; Khare et al., 2013; Lai et al., 2013; Liu et al., 2013; Marques & Rinaldi, 2013; Peng et al., 2013; Rak et al., 2013). As a result, a number of tools using BioC can be downloaded and applied to any BioC-formatted dataset as well as combined with other existing processes. Natural language processing often begins with a linguistic pre-processing pipeline. Two commonly used tool sets for biomedical text data, MedPost (Smith et al., 2004) and Stanford (Klein & Manning, 2003), have been adapted to the BioC format. The BioC NLP pipelines (Comeau et al., 2013) can perform sentence segmentation, tokenization, lemmatization, stemming, part-of-speech tagging and parsing. The BioC format also allows the convenience of mixing and matching different tools, regardless of whether the work medium is in C++, Java, or another language with BioC support. As a result, researchers in a team that have different preferences in programming languages can easily work together.

Abbreviation definition identification, which has been considered an important step in biomedical entity recognition tasks, is now available in BioC (Islamaj Dogan et al., 2013) via the implementation of three different algorithms: Schwartz and Hearst (Schwartz & Hearst, 2003), Ab3P (Sohn et al., 2008) and NatLAb (Yeganova et al., 2011). Schwartz and Hearst is a well-known, simple, and effective algorithm. Ab3P uses a richer rule-based approach with rules adapted to the length of the short form and designed to take precedence based on an approximate precision measure. NatLAb was developed as a machine learning approach, with rules as features, and was trained on a naturally labeled training set that contrasted potential definitions with random analogs to identify the long form definition for each abbreviation.

In addition, a number of biomedical named entity recognition (NER) tools were adapted to work seamlessly with the BioC format. These include a suite of tools for recognition of diseases, mutations, and chemical names, as well as gene and species normalization (Khare et al., 2013), and metabolic process concept identification (Rak et al., 2013). The results of these tools can be used directly or as features for more sophisticated entity recognition or understanding tasks.

Other contributions to BioC are tools that convert other data formats to BioC. Of note is the Brat2BioC tool, which provides two-way conversion between BRAT (brat rapid annotation tool, the BioNLP Shared Task series' data file format) and BioC (Jimeno Yepes et al., 2013). This functionality was realized by two other teams (Rak et al., 2013) and (Peng et al., 2013) as well, as part of their BioCreative IV contributions. This allows researchers to intermingle resources in either format. In addition, PubTator (Wei et al., 2013), a web-based annotation tool, has also been adapted to BioC. Finally modules that perform higher level tasks such as semantic role labelling (Lai et al., 2013) or sentence simplification (Peng et al., 2013) are available to the community via web access.

4. BioC-formatted corpora

The BioC website lists a number of biomedical corpora for download and provides links to other websites that host biomedically-relevant BioC-formatted corpora. These corpora are mostly based on PubMed, and they can all be used, alone or in combination, for the development and analysis of new biomedical language processing methods and techniques.

For example, the BioC website hosts four corpora annotated for abbreviation definition in biomedical literature, namely the Schwartz and Hearst corpus of 1000 PubMed abstracts(Schwartz & Hearst, 2003), the BIOADI corpus of 1201 PubMed abstracts (Kuo et al., 2009), the Ab3P corpus of 1250 PubMed abstracts (Sohn et al., 2008) and the old MEDSTRACT corpus of 199 PubMed abstracts. They are all manually annotated and curated and there exists virtually no overlap among them, which demonstrates that they may be usefully combined into a larger resource [Islamaj Dogan et al. submitted].

Close to twenty biomedical corpora in BioC format (Jimeno Yepes et al., 2013) may be downloaded from the WBI repository[4]. These corpora include annotations for entities such as genes, mutations, and chemicals, as well as relations such as protein-protein interactions, disease-treatment relations, gene-expression and others.

Another important contribution to the BioC repository is a corpus of 200 full-text articles, the BC4GO corpus, annotated for gene ontology (GO) terms (Khare et al., 2013). This corpus is annotated for GO terms, GO evidence codes and key sentences that annotators used as evidence for the annotations, and was the official corpus for the BioCreative IV GO task.

Finally, the NCBI disease corpus (Dogan et al., 2014), which contains annotations for disease mentions and concepts in a collection of 800 PubMed citations, manually annotated by 14 annotators, is also available in BioC format.

5. The BioC-PMC Corpus

Here we present and make available to the biomedical community the largest BioC corpus yet of biomedical literature that can facilitate biomedical text processing. The BioC-PMC (PMC Open Access articles in BioC format) corpus contains all full-text articles currently available for download at the PMC ftp site[5].

The BioC-PMC corpus currently contains 581,779 articles from more than 3000 journals. These articles are also available for download in the PMC XML format, or PDF. There are supplementary files available via the PMC ftp site, and users who are interested in complete information in the original formatting are encouraged to obtain those files.

The articles, as originally formatted for publication, contain similarities as well as quite a few differences in formatting, depending on the publisher. Full-text articles formatted to be displayed on and downloaded from the PMC website, are the end result of a considerable body of work attempting to unify the large array of article formats used by all contributing journals and publication venues. As such, while the end result is admirable and provides a unified view for all full-text articles displayed on the website, the underlying XML contains a huge amount of mark-up information that is generally not valuable for biomedical text processing.

When full-text articles are converted to BioC format, an article's textual information is preserved, and appears in passages. The BioC format for each document is organized as follows:

1. Document ID, which corresponds to the PubMed Central ID (PMCID).

2. First passage contains the front matter—which contains author names, article title (text of passage), article id (both PubMed and PMCID), and publisher information.

3. The rest of the article is a series of passages, of which the abstract is distinguished, and, if it is a

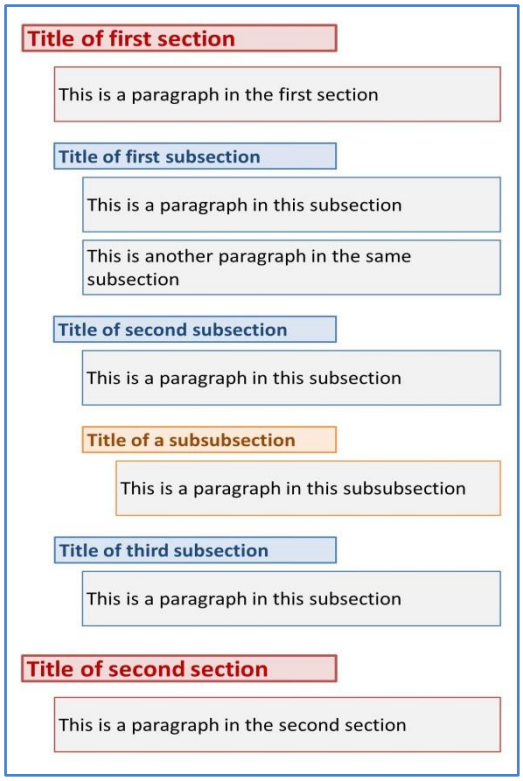


Figure 1 Thumbnail web view of an article.

[4] http://corpora.informatik.hu-berlin.de/
[5] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Passage type	Description of text in passage
title_1	First section title
paragraph	Paragraph in section
title_2	Subsection title
paragraph	Paragraph in subsection
paragraph	Another paragraph in subsection
title_2	Title of second subsection
paragraph	Paragraph in subsection
title_3	Title of a subsubsection
paragraph	Paragraph in subsubsection
title_2	Title of third subsection

Table 1: The linear structure of Figure 1 represented in BioC passages. The types allow recovery of the nested structure if desired.

structured abstract, this structure is reflected in passages. This idea is continued in the whole article, as the information about the original nested sections is preserved in the passage types.

4. Figure captions appear as separate passages, as do table captions and footnotes. However, the table structure is lost, and all text previously organized in a table now exists as a sequence of text fragments in a corresponding passage.

5. Each reference is a separate passage. The title of the referenced article is the passage text. Other details vary with the elements in the original PMC XML file.

Care has been taken that all relevant textual information in an article is preserved. While the passages appear in linear order, information about the original nested structure appears in the passage types, so the original structure can be recovered. Figure 1 and Table 1 demonstrate how the nested structure of the document can be reconstructed from a linear series of passages. First, the passages appear in a simple, linear order. If the processing cannot benefit from the structured nature of the document, it can be easily ignored. However, if the structure can inform the processing, much of it can be recovered. The title passages directly indicate section titles and the proper nesting depth of the passage. They also indirectly indicate the end of the previous section.

Details about the passage types are given in the `keyfile`. The BioC-PMC text is available in both Unicode and ASCII. The Unicode characters include all original characters in the PMC file, which occasionally stretch beyond the Basic Multilingual Plane.

We plan to continue to improve the BioC-PMC corpus files, and add additional information from the full PMC XML file, as long as it can be done consistently with the goals of BioC. One possible example is indicating where the references to figures and tables appear in text via annotations and relations. Another could be identifying bold and italic text using annotations.

5.1 BioC-PMC corpus characteristics

The current release of the BioC-PMC corpus was downloaded in July 2013 and contains 581,779 articles. Of these, about 86,000 are not fully available in XML, but exist partially as OCR'ed image files. These documents make up the two high peaks in Figure 2, because the PMC XML files of these articles contained only the title or title and abstract. The current size of the corpus is 64GB.

BioC-PMC articles have 83 passages on average. We calculated the average size of an article as the number of BioC passages. The number of BioC passages in a document is dependent on several things: 1. the authors' decisions to outline and partition their work when writing the article, 2. the suggested or required article format by publication venue, and finally 3. the rendering of that format to the PMC XML format, which in turn is simplified in BioC. This distribution is depicted in Figure

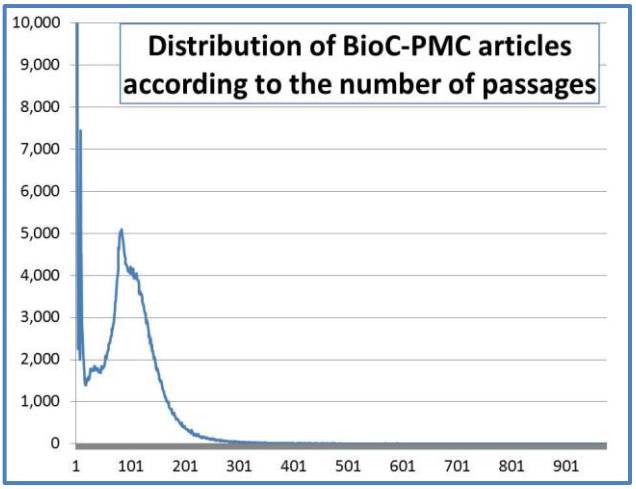


Figure 2: Number of BioC-PMC articles (y-axis) per number of passages in an article (x-axis).

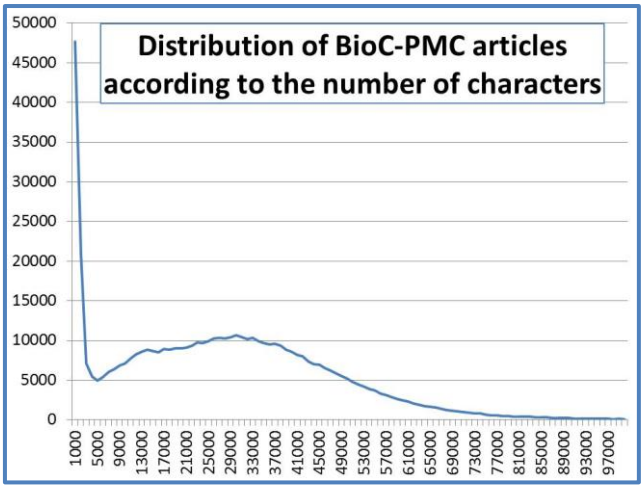


Figure 3: Number of BioC-PMC articles (y-axis) per number of characters in an article (x-axis).

Articles	Publication venue
53,075	PLoS One
22,510	The Journal of Experimental Medicine
22,372	The Journal of Cell Biology
20,894	British Journal of Cancer
20,837	Acta Crystallographica Section E: Structure Reports Online
14,010	Environmental Health Perspectives
10,333	Nucleic Acids Research
7,801	Critical Care
7,786	The Journal of General Physiology
7,240	The Yale Journal of Biology and Medicine

Table 2: Ten most frequent journals in the BioC-PMC corpus.

2, and shows a range from 1 to more than 100,000. The clear majority of articles, 97% of them, have less than 200 passages. The average and median number of passages per article are 83 and 84, respectively. Exceptions include 12 articles of more than 10,000 passages. The PMC collection includes articles which are conference transcripts, and these transcripts tend to be very long.

We also computed BioC-PMC article length in text characters, and this is shown in Figure 3. The average number of ASCII text characters per article is 26,242, with a median of 24,846. The article you are reading has a similar length. Only 5% of PMC articles have a character length of more than 100,000 and approximately 50,000 articles in the current collection have less than 1,000 text characters.

Top Ten Journals in the BioC-PMC corpus. We examined the distribution of BioC-PMC articles across their publication venues. As mentioned above, there are more than 3,000 journals and other publication venues that contribute articles to the PMC OA subset. To give an example of the range of topics covered, in Table 1, we show the number of articles for the top ten most common journals.

Section Titles in the BioC-PMC corpus. When converting the PMC-OA corpus to BioC, we were careful to preserve information about the original sections of articles. Researchers, however, may be interested in working only on particular sections of a full-text article, as different pieces of information tend to be concentrated in pre-determined sections of an article. For example Materials and Methods may need to be treated differently than the other sections. To explore this line of research, we extracted all level one headings in the BioC-PMC corpus and tried to find common section titles that appear in the majority of articles. While the expected sections of AIMRD (Abstract, Introduction, Methods, Results, Discussion) format were the most common, we ended up with a list of more than 200,000 unique level one section

Common Article sections
Introduction, Background, INTRODUCTION, The Study, Purpose, Objective, Related literature, Related Literature, Review
Methods, Materials and Methods, Materials and methods, MATERIALS AND METHODS, METHODS, RESEARCH DESIGN AND METHODS, Patients and methods, Material and Methods, Implementation, Methodology, PATIENTS AND METHODS, Methods/Design, MATERIAL AND METHODS, Subjects and Methods, Patients and Methods, Methods Summary, EXPERIMENTAL PROCEDURES
Results, RESULTS, Results and Discussion, Results and discussion, Experimental Section, Results/Discussion, Findings
Discussion, DISCUSSION, Limitations
Conclusion, Conclusions, CONCLUSION, CONCLUSIONS, Concluding remarks, Discussion and conclusion, Concluding Remarks, Discussion and Conclusions, Summary
Acknowledgements, Acknowledgments, Acknowledgement, Acknowledgements and Funding, Acknowledgements and funding, ACKNOWLEDGEMENTS, FUNDING, Funding
Authors' contributions, Author Contributions, Authors' contribution, Authors' Contributions, Author's contributions
Competing interests, Conflict of Interest Statement, Competing interest, Conflict of interest and funding, Disclosures, CONFLICT OF INTEREST, Competing Interests
Supplementary material, Supporting Information, Special details, SUPPLEMENTARY DATA, Supplementary data, Supplemental Material, Additional data files, SUPPLEMENTARY MATERIAL
Availability and requirements
Authors' information
Figures, Figures and Tables
Pre-publication history
CASE REPORT, Case presentation, Case Report, Case report, Case Presentation
Appendix
Abbreviations, List of Abbreviations, List of abbreviations used

Table 3: Most common section titles and their synonyms in BioC-PMC corpus.

titles. In Table 3, we grouped level one section titles that appeared in more than 500 articles in the BioC-PMC corpus into intuitive common groups, in order to be able to list common synonyms and identify similar sections in the data. As expected, the majority of differences are

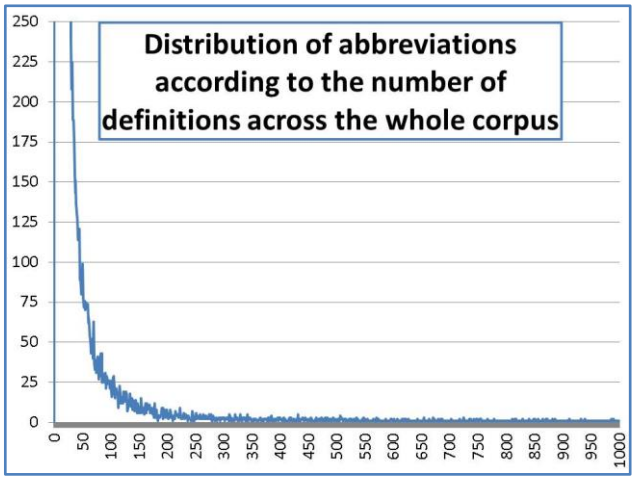


Figure 4: The distribution (from 581,779 full text articles) of the number of abbreviations paired with their different long forms. The y-axis shows the number of abbreviations, and the x-axis the number of distinct long form definitions for each abbreviation.

inflectional variants, spelling variants, lowercase versus uppercase usage, etc. Table 3 excludes section titles from the journal *Acta Crystallographica Section E: Structure Reports Online*. This journal, has contributed more than 20,000 articles to the PMC OA subset, and these articles have a very distinct format, with section titles such as: *Fractional atomic coordinates and isotropic or equivalent isotropic displacement parameters (A2),*and *Atomic displacement parameters (A2),* which are not found in articles from other journals.

6. An example application: abbreviation definition identification over BioC-PMC

To verify the usability of the BioC-PMC corpus for actual research, we performed a simple experiment of collecting abbreviation definitions in biomedical articles. We used the BioC compatible Ab3P tool to identify abbreviations and their definitions on all the text elements in the BioC-PMC corpus.

Abbreviations are consistently used both in biomedical literature and as search terms in biomedical databases. Correct resolution of an abbreviation to its intended meaning requires that the abbreviation definitions are identified as the authors introduce them in the text. We report the extent of abbreviations in full text articles, which has never been done on this scale before. Here we present our observations from this study, and the whole set of abbreviations and their definitions that we have found in this corpus is available upon request.

6.1 Abbreviations in full-text articles

We extracted all abbreviations and their definitions from the articles and where they were found and computed the statistics of Table 4.

Most abbreviation definition finding algorithms are trained on and applied to text in titles and abstracts. Table 3 reveals several interesting facts. First, the majority of abbreviations are defined in the body of an article. Out of 1,346,388 unique abbreviation definitions found in the body of the article, excluding title and abstract (a total of 5,185,160 short form, long form pair occurrences), only 11% (152,980) of those are also found in a title or abstract. This is another example of the importance of full-text processing for biomedical text mining. Second, figure and table captions seem a popular place to define abbreviations. Figure and table captions have been established as important information extraction locations for biomedical text mining (Hearst et al., 2007). Our findings reinforce this conclusion, and even more so, when we look deeper into the numbers. Of 265,641 unique abbreviation definitions extracted from figure and table captions (a total of 640,583 short form, long form pair occurrences), only 46,519 (17.5%) are found defined in a title or abstract of the BioC-PMC corpus, and only 111,489 (42%) are defined elsewhere in the BioC-PMC corpus, including title and abstract.

Article Section	Number of abbreviation definitions	Percentage
Title	14,107	0.24%
Abstract	669,120	11.40%
Table caption	83,911	1.43%
Figure caption	556,672	9.49%
References	392,812	6.69%
Text	4,151,765	70.75%
Total	5,868,387	100.00%

Table 4: Total number of abbreviation definitions in the BioC-PMC corpus as found in different sections of an article.

Unique abbreviation definitions	1,409,957
Unique abbreviations	372,161

Table 5: Abbreviation statistics. The number of unique abbreviation definitions in the whole BioC-PMC corpus and the number of unique abbreviations.

Figure 4 reports data on the abbreviation definition pairs as found across the whole corpus. Table 5 lists the number of unique definitions as found across the whole corpus, and the number of unique abbreviations. This demonstrates that abbreviations are ambiguous and often the same short form is paired with multiple different long forms. To examine the extent of this phenomenon we plotted the distribution of abbreviations per number of definitions, and show this data in Figure 4. The majority of abbreviations, 64%, have only one definition. This is not shown in Figure 4, in order to be able to scale the data. A small minority, less than 0.8% of the total abbreviations, are paired with 100 distinct definitions or more. It is important to note that, the long form definitions while often being term variations of the same concept (for example, GFP -> “green fluorescent protein” and “Green

Fluorescent Protein"), also represent different concepts (for example PI -> "propidium iodide", "phosphatidylinositol", "protease inhibitor" etc.). These statistics demonstrate the complexity of the abbreviations in biomedical text.

The extraction of abbreviation definitions from the data in the BioC-PMC corpus utilized the BioC compatible Ab3P tool. For this study, we were able to use all the text passages contained in the corpus, for both exploration and analysis purposes. This example application of a data mining task shows that the corpus is ready for use.

7. Conclusions

BioC is a useful format and a growing set of tools and corpora that is simple, easy-to-use and freely available to the community. The most recent addition is the BioC-PMC corpus. The BioC-PMC corpus simplifies the processing of the PMC OA subset of articles for text mining purposes and our initial processing shows that it is ready for use. BioC code and data are available for download at the BioC site: http://bioc.sourceforge.net.

8. Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

9. References

Arighi, C.N., Lu, Z., Krallinger, M., Cohen, K.B., Wilbur, W.J., Valencia, A., et al. (2011). Overview of the BioCreative III Workshop. *BMC bioinformatics,* 12 Suppl 8, S1.

Blaschke, C. & Valencia, A. (2001). Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. *Comp Funct Genomics,* 2(4), 196-206.

Cohen, A.M. & Hersh, W.R. (2005). A survey of current work in biomedical text mining. *Brief Bioinform,* 6(1), 57-71.

Comeau, D.C., Liu, H., Islamaj Dogan, R. & Wilbur, W.J. (2013). Natural Language Processing Pipelines to Annotate BioC Collections with an Application to the NCBI Disease Corpus. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, 38-45.

Corney, D.P., Buxton, B.F., Langdon, W.B. & Jones, D.T. (2004). BioRAT: extracting biological information from full-length papers. *Bioinformatics,* 20(17), 3206-3213.

Divoli, A., Wooldridge, M.A. & Hearst, M.A. (2010). Full text and figure display improves bioscience literature search. *PLoS One,* 5(4), e9619.

Dogan, R.I., Leaman, R. & Lu, Z. (2014). NCBI disease corpus: A resource for disease name recognition and concept normalization. *J Biomed Inform*.

Hearst, M.A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M.A., et al. (2007). BioText Search Engine: beyond abstract search. *Bioinformatics,* 23(16), 2196-2197.

Hirschman, L., Burns, G.A., Krallinger, M., Arighi, C., Cohen, K.B., Valencia, A., et al. (2012). Text mining for the biocuration workflow. *Database (Oxford),* 2012, bas020.

Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC bioinformatics,* 6 Suppl 1, S1.

Islamaj Dogan, R., Comeau, D.C., Yeganova, L. & Wilbur, J. (2013). Finding abbreviations in biomedical literature: Three BioC-compatible modules and three BioC formatted corpora. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop* 23-30.

Jimeno Yepes, A., Neves, M. & Verspoor, K. (2013). Brat2BioC: conversion tool between brat and BioC. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, 46-53.

Khare, R., Wei, C.-H., Mao, Y., Leaman, R. & Lu, Z. (2013). Improving Interoperability of Text Mining Tools with BioC. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, 10-22.

Klein, D. & Manning, C.D. (2003). *Accurate unlexicalized parsing*. Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1.

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., et al. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol,* 9 Suppl 2, S1.

Kuo, C.J., Ling, M.H., Lin, K.T. & Hsu, C.N. (2009). BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature. *BMC bioinformatics,* 10 Suppl 15, S7.

Lai, P.-T., Dai, H.J., Wu, J.C.-Y. & Tsai, R.T.-H. (2013). A Biomedical Semantic Role Labeling BioC Module for BioCreative IV. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, 54-60.

Liu, W., Comeau, D.C., Islamaj Dogan, R. & Wilbur, W.J. (2013). Extending BioC Implementation to More Languages. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, 31-37.

Marques, H. & Rinaldi, F. (2013). PyBioC: a python implementation of the BioC core. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, 2-4.

Peng, Y., Tudor, C.O., Torii, M., Wu, C.H. & Vijay-Shanker, K. (2013). Enhancing the Interoperability of iSimp by Using the BioC Format. In *Proceedings of the Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*, 5-9.

Rak, R., Batista-Navarro, R., Rowley, A., Miwa, M., Carter, J. & Ananiadou, S. (2013). NaCTeM's BioC Modules and Resources for BioCreative IV. In *Proceedings of the Proceedings of the Fourth*

BioCreative Challenge Evaluation Workshop, 61-67.

Schwartz, A.S. & Hearst, M.A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pac Symp Biocomput*, 451-462.

Shah, P.K., Perez-Iratxeta, C., Bork, P. & Andrade, M.A. (2003). Information extraction from full text scientific articles: where are the keywords? *BMC bioinformatics,* 4, 20.

Smith, L., Rindflesch, T. & Wilbur, W.J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics,* 20(14), 2320-2321.

Sohn, S., Comeau, D.C., Kim, W. & Wilbur, W.J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics,* 9, 402.

Wei, C.H., Kao, H.Y. & Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res,* 41(Web Server issue), W518-522.

Wu, C.H., Arighi, C.N., Cohen, K.B., Hirschman, L., Krallinger, M., Lu, Z., et al. (2012). BioCreative-2012 virtual issue. *Database (Oxford),* 2012, bas049.

Yeganova, L., Comeau, D.C. & Wilbur, W.J. (2011). Machine learning with naturally labeled data for identifying abbreviation definitions. *BMC Bioinformatics,* 12 Suppl 3, S6.

Yu, H. & Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics,* 19 Suppl 1, i340-349.

Annotating Question Types for Consumer Health Questions

Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman

National Library of Medicine
National Institutes of Health
Bethesda MD, USA
kirk.roberts@nih.gov; ddemner@mail.nih.gov

Abstract

This paper presents a question classification scheme and a corresponding annotated corpus of consumer health questions. While most medical question classification methods have targeted medical professionals, the 13 question types we present are targeted toward disease questions posed by consumers. The corpus consists of 1,467 consumer-generated requests for disease information, containing a total of 2,937 questions. The goal of question type classification is to indicate the best strategy for selecting answers from publicly available health information resources, such as medical encyclopedias and medical research websites. The annotated question corpus, along with detailed annotation guidelines, are publicly available.

Keywords: question classification, question answering, medical language processing

1. Introduction

Most research in automatic question answering has focused on well-formed factoid questions in the style of the TREC Question Answering evaluations (Prager, 2006). However, real-life questions that automatic systems must handle are typically not well-formed or explicit in their information needs. This is particularly the case for consumer health questions. Zhang (2010) focuses on consumer health questions submitted to Yahoo Answers, finding these questions were largely concerned with diseases and symptoms, contained many abbreviations and misspellings, and often contained more than one actual question.

Furthermore, the type of answer to be provided to a consumer is qualitatively different than an appropriate answer for a medical professional. Many questions that the general public might have about diseases are answerable by reliable consumer health information resources, such as MedlinePlus[1] or Genetics Home Reference[2]. Even these consumer-oriented resources might be difficult to navigate for users with little medical knowledge. Thus, a question answering system that could point a user to the most relevant encyclopedic articles or sections could prove useful. Some information providers, such as the NIH Genetic and Rare Diseases Information Center[3] (GARD) not only answer consumers' questions, but also make the questions and the answers publicly available.

In our own consumer question answering system, we have found a wide variety in the styles and types of questions even in a relatively focused area of questions about specific diseases, especially their treatments and outlook. Our rule-based classification of these questions into traditional treatment, cause, prognosis, diagnosis, manifestation, and general information types is sufficient for understanding and structuring simple questions, but it is insufficient for most of the questions submitted to our system. It is also clear that it is practically impossible to develop rules for understanding the many types of complex questions, and thus a corpus-based machine learning solution is applicable. To our knowledge, there are few publicly available collections of consumer health questions, and none annotated at the granularity needed for our task. GARD questions are a good intermediate step from the simple questions, such as "*is there any help availabe,for fibromalgia*" that our system translates to "*What are treatments for fibromyalgia?*" to complex questions, such as the one in Figure 1.

> My wife has been having shortness of breath in the mornings (mornings only). She has, what I think, excessive heart rate as well. Could this be anxiety attacks? I don't think it would be a heart issue. She certainly isn't overweight. SHe is 5'2"; and 100-105 pounds. if she is having anxiety attack... what is the best course of action? She seems to feel better when she lies down and rests.....while either watching TV...or sleeps. SHe has no history in her family for heart disease either. WHat are your thoughts?

Figure 1: Consumer health question.

GARD questions, for example in Figure 2, are well-formed with few misspellings and are primarily focused on a specific condition. Questions in the GARD corpus are obtained from the public and answered by experienced information specialists. We have annotated these questions with more detailed question types as the next step towards understanding a wide range of consumer health questions.

> Do children with Cat Eye Syndrome generally experience a decline in physical abilities as they reach adulthood? Is there any shortening of the lifespan associated with this condition?

Figure 2: Consumer health question from GARD.

For clarity, the remainder of this paper refers to the multi-sentence, possibly multi-question spans found in Figures 1 and 2 as *requests*. A *question* refers to a clause asking for a single piece of information. Section 2 of this paper describes previous work in question classification; Section 3 briefly outlines our proposed question types; Section 4 describes the process of annotating the GARD corpus; Section 5 provides numerous examples of each question type; and Section 6 describes our corpus.

[1] http://www.nlm.nih.gov/medlineplus/
[2] http://ghr.nlm.nih.gov/
[3] https://rarediseases.info.nih.gov/GARD/

2. Background

In much of the literature, question classification is synonymous with answer type detection (Hermjakob, 2001). An answer type corresponds to the answer's expected class, typically a named entity type for TREC QA style factoid questions. For instance "*Who is Tom Cruise married to?*" has a PERSON answer type, while "*What was the capital of the Holy Roman Empire?*" has a CITY answer type. Li and Roth (2002) provide one of the first statistical methods for classifying answer types, along with a corpus on which many other methods have been judged. Their answer types form a 2-level hierarchy with six "coarse" types and fifty "fine" types. Roberts and Hickl (2008) demonstrate how such an answer type hierarchy can be arbitrarily extended as the need arises. The automatic classification of answer types has been shown to be a highly semantic task (Metzler and Croft, 2005) due to the short length and predictable syntactic structures of TREC-style questions. The single most important semantic feature involves finding and typing the answer type term (*capital* in the above question) to the answer type (Krishnan et al., 2005; Hickl et al., 2007).

Despite the focus on answer type detection, there are other important forms of question classification, especially when question answering systems are designed to answer non-factoid questions or questions that require domain knowledge. For instance, the question "*List the cities of the Holy Roman Empire.*" still has a CITY answer type, but also requires classifying the question's function (Bu et al., 2010). Additionally, it was useful to identify a question's topic or domain (Duan et al., 2008; Roberts and Harabagiu, 2012; Chan et al., 2013), especially when further domain knowledge is necessary to answer the question. In addition to the question classification tasks listed above, innumerable other tasks exist for the many classes of questions found in natural language (Chali and Hasan, 2012; Banerjee and Bandyopadhyay, 2012).

In contrast to the forms of question classification in TREC QA style questions, medical questions have presented very different classification methods. Ely and colleagues (1999; 2000; 2005) collected several sets of medical questions posed by physicians. These questions are available as part of the NLM Clinical Questions Collection[4] (CQ). On these well-studied question sets, automatic approaches have been proposed to classify question answerability (Yu and Sable, 2005), taxonomy (Kobayashi and Shyu, 2006), and topic (Yu and Cao, 2008). The diversity of possible ways medical questions may be classified speaks to their potential complexity. Because of this complexity, medical question answering does not appear to be a promising field for purely factoid question answering approaches. For this reason, many medical question answering systems utilize the PICO structure (Demner-Fushman and Lin, 2007; Schardt et al., 2007), which encodes a far wider range of information-seeking queries.

The complexity of medical questions is further compounded when answering questions posed by consumers, as shown in Figures 1 and 2. While answer types are certainly important in returning appropriate medical information to the consumer, we have found it is more appropriate to determine the general type of information the user is looking for (treatment, prognosis, symptoms, etc.). By classifying questions along these lines, then, our question answering system may choose the most appropriate passage selection strategy.

In this paper we present a set of question types and a corresponding annotated corpus designed for selecting an appropriate strategy for consumer health questions. While previous work has demonstrated the value of deep question taxonomies, we have focused only on an appropriate first-level set of types for now. The specific impact of dealing with consumer health questions (as opposed to questions posed by medical professionals) is two-fold. First, certain ambiguities often present in consumer questions result in ontological choices about how best to organize the first level of our type scheme. Second, the distribution of question types can differ greatly between professional and consumer questions. While medical question answering systems for professionals often focus on the PICO structure, consumer requests often contain more questions for general information, causes, and prognosis (Liu et al., 2011; Kilicoglu et al., 2013). Our types thus focus on classifying the most common classes of consumer questions, and ignore the PICO structure. While the corpus we present focuses on genetic and rare diseases, we believe the following question types generalize well to other disease questions posed by consumers.

3. Question Types

We propose the following question types. Examples and annotation rules are provided in Section 5. As an example of how the question types correspond to encyclopedic sections, for each question type we list the sections of MedlinePlus where answers are most likely to be found (though our system uses several other sources as well).

(1) ANATOMY: Identifies questions asking about a particular part of the body, such as the location affected by a disease. Answers to ANATOMY questions are typically found in the "Anatomy/Physiology" section.

(2) CAUSE: Identifies questions asking about the cause of a disease. This includes both direct and indirect causes, such as factors that might increase the risk of developing a disease. Answers to CAUSE questions are typically found in the "Common Causes" section.

(3) COMPLICATION: Identifies questions asking about the problems a particular disease causes. This primarily focuses on the risks faced by patients with the disease and does not include the signs/symptoms of a disease. Answers to COMPLICATION questions are typically found in the "Related Issues" section.

(4) DIAGNOSIS: Identifies questions asking for help making a diagnosis. Our question answering system is not designed to provide a direct diagnosis. However, the DIAGNOSIS type does include questions asking about diagnostic tests, or methods for determining the difference between possible diagnoses (differential diagnosis). Answers to DIAGNOSIS questions are typically found in the "Diagnosis" and "Testing" sections.

[4] http://clinques.nlm.nih.gov/

(5) INFORMATION: Identifies questions asking for general information about a disease. This includes type information about diseases (e.g., whether two disease names represent the same disease, or if one disease is a type of another disease). Answers to INFORMATION questions are typically found in the "Definition" and "Description" sections.

(6) MANAGEMENT: Identifies questions asking about the management, treatment, cure, or prevention of a disease. Answers to MANAGEMENT questions are typically found in the "Treatment" and "Prevention" sections.

(7) MANIFESTATION: Identifies questions asking about signs or symptoms of a disease. Answers to MANIFESTATION questions are typically found in the "Symptoms" section.

(8) OTHEREFFECT: Identifies questions asking about the effects of a disease, excluding signs/symptoms (MANIFESTATION) or predispositions (COMPLICATION). When the question requires medical knowledge to understand a given effect is actually a MANIFESTATION or COMPLICATION, it is instead classified as an OTHEREFFECT, the reasoning for which is provided in Section 4. Answers to OTHEREFFECT questions are typically found in the "Symptoms" and "Related Issues" sections.

(9) PERSONORG: Identifies any question asking for a person or organization involved with a disease. This can include medical specialists, hospitals, research teams, or support groups for a particular disease. Answers to PERSONORG questions are typically not found in MedlinePlus articles, but may be found in links on MedlinePlus pages (especially for support groups).

(10) PROGNOSIS: Identifies questions asking about life expectancy, quality of life, or the probability of success of a given treatment. Answers to PROGNOSIS questions are typically found in the "Expectations" section.

(11) SUSCEPTIBILITY: Identifies questions asking how a disease is spread or distributed in a population. This includes inheritance patterns for genetic diseases and transmission patterns for infectious diseases. Answers to SUSCEPTIBILITY questions are typically found in the "Mode of Inheritance" and "Prevalence" sections.

(12) OTHER: Identifies disease questions that do not belong to the above types. This includes non-medical questions about a disease, such as requests for financial assistance or the history of a disease. Answers to OTHER questions are typically found in the "Definition" and "Description" sections, though by definition these answers may be found anywhere in the encyclopedic entry.

(13) NOTDISEASE: Identifies questions that aren't about a disease and thus not handled by our question answering system.

Additionally, we propose one question attribute that could apply to any question of the above types:

(14) RESEARCH: Identifies questions asking for the latest research, such as medical publications or clinical trials. Answers to RESEARCH questions are typically found using specialized search engines such as `PubMed`[5] for medical publications or `ClinicalTrials.gov`[6] for clinical trials.

The choice of these question types started with an initial set (Van Der Volgen et al., 2013), but was altered over the course of the annotation process described in the next section. The final set of types is thus both based on *a priori* medical knowledge and specific issues that arise in consumer health questions.

To a large extent, the types are generalizable beyond consumer health questions, though the choice of how abstract or how specific the types should be was based on their distribution in consumer health questions. For example, the types overlap quite a bit with those in the CQ set. In some cases, the CQ types are more broad than our types, and in other cases they are more specific.

4. Annotation Process

As a source for consumer health questions about diseases, we used 1,467 publicly available requests on the GARD website. Because these requests contain multiple question sentences, and many of the question sentences contain requests for different types of information, the questions were annotated for syntactic decomposition as described in Roberts et al. (2014). Figure 3 illustrates an example of question decomposition. This decomposition process results in 2,937 questions from the 1,467 GARD requests.

Request:
I have been recently diagnosed with antisynthetase syndrome. Could you please provide me with information on antisynthetase syndrome? I am also interested in learning about prognosis, treatment, and clinical trials.
Decomposed Questions:
Q1) Could you please provide me with information on antisynthetase syndrome?
Q2) I am also interested in learning about prognosis.
Q3) I am also interested in learning about treatment.
Q4) I am also interested in learning about clinical trials.

Figure 3: Question Decomposition Example.

Additionally, these requests are annotated with a FOCUS disease, typically one per request but occasionally more than one. The FOCUS is the disease the consumer is interested in learning more about. In Figure 2, the FOCUS is *Cat Eye Syndrome*. In Figure 3, the FOCUS is *antisynthetase syndrome*. The FOCUS is not only important as a default value for resolving co-reference, but it is also useful for determining the question type. Since our question answering approach primarily utilizes medical encyclopedias, the FOCUS is often the encyclopedic entry, while the question type is the section of the entry in which the answer is found. For cause-and-effect questions, for example, if the FOCUS is the cause, the answer is likely to be found in the effect section, and vice versa.

[5] http://www.ncbi.nlm.nih.gov/pubmed

[6] http://clinicaltrials.gov/

A pool of five annotators (all the authors) were used: two MDs (MF, DDF), three computer science PhDs (KR, HK, DDF), and one medical librarian (KM). Annotation was performed with a custom annotation tool. On average, annotators were able to manually classify approximately 10 questions per minute.

The initial set of question types was based on the study of consumer health questions by (Van Der Volgen et al., 2013). The eleven question types in that study were: ANATOMY, CAUSE, COMPLICATION, DIAGNOSIS, INFORMATION, LOCATION, MANIFESTATION, PREVENTION, PROGNOSIS, SUSCEPTIBILITY, and TREATMENT. Using these types as well as the OTHER type as a starting point, three annotators double-annotated the questions from 250 GARD requests. The average inter-annotator agreement (*Kappa*) for this round was 0.802, which is reasonably good agreement. To resolve annotation conflicts, all five annotators met in person to discuss major sources of disagreements. During the resolution process, it was decided that three changes would be made. First, the RESEARCH attribute was added as it was felt this was an important distinction that indicated consumer-oriented medical encyclopedias were not sufficiently detailed to answer these questions. Second, the LOCATION type was re-named PERSONORG to emphasize the goal of looking for medical professionals or organizations. Previously, requests for websites had been considered LOCATIONs, but a website can contain any type of information, and would thus only be considered an answer type as opposed to a search strategy. Finally, an ASSOCIATION type was added when a question asked about the relationship between multiple diseases.

In the second round, four annotators (two of whom participated in the previous round) double-annotated questions from 450 GARD requests. The inter-annotator agreement for this round was a slightly lower 0.742. This drop can likely be explained by the two additional annotators, but is more likely the result of adding an additional question type as both new annotators had participated in the first round conflict resolution session. During the second resolution process, it was decided to make two additional changes. First, the NOTDISEASE type was split off from OTHER to indicate questions not relevant to a disease question answering system. Second, and more importantly, the OTHEREFFECT type was created. This type subsumed most of the ASSOCIATION questions, where the association is an effect of the FOCUS. All other ASSOCIATION questions involve one disease being the same as, different from, or a sub-type of another disease, and were re-classified as INFORMATION. More notably, though, OTHEREFFECT sits in the ambiguous zone between MANIFESTATION and COMPLICATION. It is difficult for an automatic system to encode sufficient medical knowledge to accurately differentiate between these two types. Indeed, this was a major source of annotator disagreement, as both are considered effects of a disease. Instead of merging both types into a single effect type, MANIFESTATION and COMPLICATION remain but require explicit lexical evidence to classify them (such as reference to *symptoms* or *risks*). When medical knowledge is needed, questions are considered OTHEREFFECT. The strategy for answering OTHEREFFECT questions is then simply the union of the MANIFESTATION and COMPLICATION strategies.

Question Type	Accuracy	κ
ANATOMY	99.6	0.314
CAUSE	98.2	0.784
COMPLICATION	97.4	0.506
DIAGNOSIS	97.8	0.843
INFORMATION	94.3	0.819
MANAGEMENT	98.4	0.925
MANIFESTATION	96.5	0.714
OTHEREFFECT	96.0	0.570
PERSONORG	98.4	0.809
PROGNOSIS	96.2	0.795
SUSCEPTIBILITY	95.5	0.804
OTHER	96.2	0.327
NOTDISEASE	99.6	0.233
Overall	81.0	0.787

Table 1: Inter-annotator agreement broken down by question type. Accuracy = % agreement between annotators, κ = Cohen's Kappa.

In the final round, all five annotators double-annotated the remaining 767 GARD requests. The inter-annotator agreement for this round was 0.791, an improvement from the previous round. No further changes were made to the annotation standard. Inter-annotator agreement for all rounds is broken down by question type in Table 1. In general, the more frequent types (see Table 2) have higher agreement scores, an exception being OTHEREFFECT questions.

5. Examples

To illustrate the question types presented in Section 3, we present the following examples from the annotated GARD data. For some types, informal sub-classes are described. At this point, however, only the 13 types are annotated.

ANATOMY questions are concerned with the body location of a disease, including where symptoms manifest (example 1) and if specific anatomical locations are possible body sites for a disease (2 & 3):

(1) *Does IP affect any areas inside the body, such as internal organs?*
(2) *I was researching metachondromatosis - is this something that when felt is attached to the bone itself?*
(3) *Can this cancer occur on someone's back?*

CAUSE questions are concerned with the cause of the FOCUS disease, including direct causes (1-6) as well as factors that might increase the susceptibility of a disease (7-10):

(1) *What causes this condition?*
(2) *Can pregnancy trigger such an issue?*
(3) *Could this be caused by hip dysplasia?*
(4) *Can in vitro fertilization cause Duane syndrome?*
(5) *Can you provide me with more information about this condition, including its causes?*
(6) *What are the genes involved?*
(7) *Are there any exogenous (outside) factors that cause Batten disease?*
(8) *Can a father taking an antidepressant contribute to the reason a baby has a trisomy?*

(9) *What increases the risk of males to have nondisjunction during meiosis?*
(10) *Has any research been done to check whether there is any viral involvement?*

COMPLICATION questions are concerned with longer term effects of a disease, usually in the form of risks (1-6), statements of the FOCUS causing some other longer-term disease (7 & 8), or direct queries of complications (9 & 10):

(1) *Are carriers of cystic fibrosis at a higher risk for other health conditions?*
(2) *Is there an increased risk of getting malignant disorders?*
(3) *Does propionic acidemia carry risk factors for hearing loss, especially progressive hearing loss?*
(4) *As an individual with G6PD deficiency, would living in a malaria risk country pose a greater risk to me compared with an individual who does not have this deficiency?*
(5) *Can you tell me if the risk for pre-eclampsia in pregnancy is increased significantly for a woman over the age of 35?*
(6) *Would my diagnosis of nail patella syndrome increase my risk for a miscarriage?*
(7) *Can Charcot-Marie-Tooth disease cause a paralyzed laryngeal nerve?*
(8) *Can this lead to a pre-cancerous condition?*
(9) *I would like to ask about the serious complications associated with this disease.*
(10) *What complications may occur in the future?*

DIAGNOSIS questions are concerned with the method of diagnosing a particular disease. These questions can simply ask about the diagnostic process for a given disease (1-6), specific types of diagnostic tests (7-9) and even questions asking for a specific diagnosis[7] (10):

(1) *Is it true that Smith-Magenis syndrome become more difficult to diagnose through genetic testing if the child is older than 24 months?*
(2) *Can Polycystic Kidney Diseases be diagnosed in pregnancy?*
(3) *Can Bloom syndrome be detected before symptoms appear?*
(4) *Can you provide me with information regarding diagnosis?*
(5) *How is it diagnosed?*
(6) *I would like to know what test needs to be performed in order to be able to rule it out.*
(7) *Could you please let me know the name of the blood test used to diagnose a vitiligo carrier?*
(8) *Is genetic testing available for Asperger syndrome?*
(9) *How can I learn more about genetic testing for this syndrome?*
(10) *How can we know if it is hereditary amyloidosis or not?*

[7]Note that our QA system cannot provide a diagnosis, so for these types of questions we simply provide a disclaimer along with general diagnostic information.

INFORMATION questions are concerned with basic information about a disease. This includes specific requests for general information (1-5) and sub-type/super-type information (6 & 7), which commonly concerns cancer (8-10):

(1) *What is this condition?*
(2) *Can you please provide me with general information about hyper IgD syndrome?*
(3) *I have osteopoikilosis and would like more information on this condition.*
(4) *How can I find information specific to adults with this condition?*
(5) *I would like general information on mucopolysaccharidosis I (MPS I).*
(6) *Is Machado Joseph disease a form of Parkinson's disease?*
(7) *What are the different types of syringomyelia?*
(8) *Is TMT a cancerous classification?*
(9) *Is someone diagnosed with teratoma with malignant transformation (TMT) considered to have cancer?*
(10) *Is astroblastoma a brain cancer?*

MANAGEMENT questions are concerned with the management, treatment, and prevention of a disease. These questions can be about general disease management (1), specific types of management (2 & 3), treatment (4-7), cure (8 & 9), and prevention information (10):

(1) *What can we do to help him?*
(2) *Is there a diet guide that can help me plan a diet that is low in protein?*
(3) *She had scalded skin syndrome when she was a baby: will she need pain relief?*
(4) *Are there new therapies for treatment of pili torti?*
(5) *What treatments are available for this condition?*
(6) *Does this mean that if I were to get cancer in the future I would not be able to be treated with chemotherapy?*
(7) *How long does it usually take for this medication to take effect?*
(8) *Is there a cure for HTLV-1 in general?*
(9) *Can the condition be cured?*
(10) *Is there anyway to prevent LHON with certain vitamins?*

MANIFESTATION questions are concerned with the signs and symptoms of a disease, including asking what the symptoms of a disease are (1-6), whether specific issues are symptoms of the disease (7-9), and the typical magnitude of the symptoms (10):

(1) *What are symptoms of ADPKD?*
(2) *I'd like to learn more about this syndrome including its symptoms.*
(3) *Are there any physical characteristics associated with the disorder?*
(4) *What symptoms may people experience as their disease progresses?*
(5) *Can women have symptoms of glucose 6 phosphate dehydrogenase (G6PD) deficiency?*
(6) *Are there any things I should look out for?*
(7) *Is fatigue a common symptom of polyglucosan body disease?*

(8) *Are severe headaches symptoms of this condition?*
(9) *Could this indicate that I am developing scleroderma inside my organs?*
(10) *In autosomal dominant polycystic kidney disease (ADPKD), does the severity of symptoms of tend to be the same among affected family members?*

OTHEREFFECT questions are concerned with the effects of a disease that are not explicitly COMPLICATIONS or MANIFESTATIONS. This can include questions about a causal link between the FOCUS and another condition (1-5) or a general link (6 & 7), questions about specific effects of a disease (8 & 9), and unspecified effects that may alter normal life processes (10):

(1) *Does Klinefelter syndrome cause weight gain?*
(2) *Does excess growth hormone have an effect on the uterus?*
(3) *Is it normal for her to be hungry all the time?*
(4) *Can hyperphenylalaninemia caused by tetrahydrobiopterin (BH4) deficiency cause problems with eating solids in babies?*
(5) *Can tuberous sclerosis affect eye closure?*
(6) *Could this be connected to the septo-optic dysplasia?*
(7) *I was wondering if people with Pfeiffer syndrome have circulatory problems in their legs.*
(8) *Is hearing loss in people with Waardenburg syndrome type II likely to be progressive?*
(9) *How common is it for only one leg to have bowing?*
(10) *Is it safe for them to bear children?*

PERSONORG questions are concerned with finding individuals or organizations that specialize in a particular disease. This typically includes individual specialists (1-5), researchers (6 & 7), and support groups (8-10).

(1) *How can I find a physician who is knowledgeable about this condition?*
(2) *Are there doctors who specialize in this condition?*
(3) *I would like information about which doctors could treat me.*
(4) *I have many questions, who can I talk to?*
(5) *Are there any doctors who specialize in this type of neurological dysfunction?*
(6) *Are there any new research studies enrolling people with carbamoyl phosphate synthase 1 deficiency?*
(7) *How can I be seen by researchers at the National Institutes of Health?*
(8) *Are there any support groups for adults with blepharophimosis type 1?*
(9) *How can I find other families that are in a similar situation?*
(10) *I have prolidase deficiency and would like contact other sufferers.*

PROGNOSIS questions are concerned with life expectancy and long-term quality of life. Specific types of questions include general prognosis (1 & 2), life expectancy (3), survival probability (4 & 5), lifestyle changes (6), long-term prospects of symptoms (7 & 8), and the long-term probability of symptoms emerging or returning (9 & 10):

(1) *What can I expect from this condition?*
(2) *What is the prognosis?*
(3) *I would like to know what the life expectancy is for people with this syndrome.*
(4) *What are her chances of survival if it is a cancerous tumor?*
(5) *Is Devic disease fatal?*
(6) *Can you do any other types of exercises if you have Wolff-Parkinson-White syndrome?*
(7) *Is there any hope for a return of smell for someone who has never smelled?*
(8) *Am I going to have this for the rest of my life?*
(9) *Is there any data that would suggest that kidney problems could show up later?*
(10) *As Burkitt is so rare, there is a lack of research whether the treatment is curative or palliative and I want to know what the chances are that it will return.*

SUSCEPTIBILITY questions are concerned with the spread and prevalence of a disease. Common types of questions include whether a disease is genetic (1 & 2), its inheritance patterns (3 & 4), whether it is considered rare (5 & 6), and its prevalence in certain populations (7-10):

(1) *Is this is a genetic disease?*
(2) *Is idiopathic thrombocytopenic purpura hereditary?*
(3) *Is there any evidence that hemorrhagic shock and encephalopathy syndrome is passed from parent to child?*
(4) *Will our baby inherit this condition from him?*
(5) *How rare is this disorder?*
(6) *Is Gilles de la Tourette syndrome a rare disease?*
(7) *Are people of certain religious backgrounds more likely to develop this syndrome?*
(8) *If so, what are the chances for an individual to have both conditions?*
(9) *Could you tell me which database I could look at to find the most recent data on neuroblastoma epidemiology; in particular the prevalence in Europe?*
(10) *How many cases of this condition have been identified throughout the world?*

OTHER questions are by definition outliers from the normal question types our QA system is concerned with, though they are about a specific disease. Some more common include questions about a disease's name or history (1-3) and financial considerations (4-6), though many more types of OTHER questions could be categorized (7-10):

(1) *How did Zellweger Syndrome get its name?*
(2) *Does the "M" in MOPD stand for microcephaly or Majewski?*
(3) *How will the recent discovery of the gene related to Kabuki syndrome affect her?*
(4) *Does Medicaid cover genetic testing?*
(5) *Are there any programs in New York to help with the cost of his dental work?*
(6) *Is there any funding out there to help me start some research on the effects of cooking naturally for children with this condition?*
(7) *Can I donate blood?*
(8) *Can any doctor cure Buschke Lowenstein tumor?*

(9) *Does this type of infection get reported to the department of health?*

(10) *Could my family, which has French Canadian heritage, be related to the Canadian family described by Renpenning in the 1960s?*

NOTDISEASE questions in the context of the GARD corpus are generally about genetic tests (1), treatments (2), anatomical objects (3), or gene functions (4 & 5). While there are question types related to many of these (e.g., DIAGNOSIS, MANAGEMENT, ANATOMY), these questions are not specifically about a disease and thus cannot be answered by a disease question answering system:

(1) *Are there factors that may contribute to a falsely positive result?*

(2) *However, being a dextrocardia patient, I wonder if a pacemaker can be put in my heart.*

(3) *Is complex III in the mitochondria itself or is it part of the body?*

(4) *Is there any information about what those genes do?*

(5) *Is there any information about what genes are on chromosome 20q12?*

RESEARCH questions are concerned with the most recent scientific information, and thus the answers likely wouldn't be found in a medical encyclopedia, but in medical publications or clinical trial databases. As the examples below illustrate, RESEARCH questions may come from almost any question type, including CAUSE (1), DIAGNOSIS (2), INFORMATION (3 & 4), MANAGEMENT (5 & 6), OTHEREFFECT (7), PERSONORG (8), PROGNOSIS (9), and SUSCEPTIBILITY (10):

(1) *Has any research been done to check whether there is any viral involvement?*

(2) *If cystic fibrosis testing was done by amniocentesis 6 years ago, have there been any new findings since then?*

(3) *Are there any clinical studies currently being performed?*

(4) *Could you please provide some of the latest information on this disease?*

(5) *Are there any new drugs (even under trial) to treat this condition.*

(6) *I would like to know whether current research is finding ways to cure this disease.*

(7) *Is there any research being done regarding a possible association between a history of diabetes and a diagnosis of Ledderhose disease?*

(8) *Is there any research studies enrolling women with this tumor?*

(9) *Where can I find studies which discuss these risks?*

(10) *Could you tell me which scientific reference I could look at to find the most recent data on neuroblastoma epidemiology; in particular the prevalence in Europe?*

6. Corpus Description

The annotated corpus contains 1,467 requests, with a total of 2,937 decomposed questions. The frequencies of the question types in this corpus is shown in Table 2. The most common question types are MANAGEMENT and INFORMATION, with almost one-third of all requests containing a question of this type. The next most common types are SUSCEPTIBILITY, OTHEREFFECT, PROGNOSIS, and DIAGNOSIS.

Question Type	# Questions	# Requests with Type
ANATOMY	12	7
CAUSE	128	106
COMPLICATION	38	34
DIAGNOSIS	240	181
INFORMATION	531	445
MANAGEMENT	683	497
MANIFESTATION	104	85
OTHEREFFECT	277	199
PERSONORG	126	107
PROGNOSIS	311	244
SUSCEPTIBILITY	422	324
OTHER	52	40
NOTDISEASE	13	9

Table 2: Frequencies of question types in GARD data as well as the frequencies of requests with at least one question of a given type.

Table 3 shows the top ten words for each question type as ranked by the Fisher Exact Test. For classes with a significant number of annotations (e.g., MANAGEMENT, DIAGNOSIS), these words largely conform to our intuitions. Many of these words were used as trigger words by our previous, rule-based classification system. The fact that they are highly associated in the corpus means that a machine learning system trained on this data should be able to accurately identify questions of these types.

The annotated data set is available on our project website.[8]

7. Conclusion

We have presented a question classification scheme and an annotated corpus based on 2,937 consumer health questions. An automated system trained on this data should be able to aid a consumer question answering system find appropriate strategies for retrieving health information. In future work, we plan to experiment with machine learning approaches to this task as well as determine how well it generalizes to other, more grammatically challenging datasets.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. We would additionally like to thank Stephanie M. Morrison and Janine Lewis for their help accessing the GARD data.

8. References

Somnath Banerjee and Sivaji Bandyopadhyay. 2012. Question Classification and Answering from Procedural Text in English. In *COLING Workshop on Question Answering for Complex Domains*, pages 11–26.

[8] http://lhncbc.nlm.nih.gov/project/consumer-health-question-answering

Question Type	Top 10 Words
ANATOMY	*areas*, *inside*, *ip*, *body*, *affect*, *attached*, *felt*, *metachondromatosis*, *internal*, *researching*
CAUSE	*causes*, *cause*, *caused*, *taking*, *exposure*, *factors*, *trisomy*, *genes*, *chemical*, *increases*
COMPLICATION	*risk*, *complications*, *pre*, *individual*, *result*, *compared*, *country*, *greater*, *malaria*, *paralyzed*
DIAGNOSIS	*testing*, *diagnosed*, *diagnosis*, *test*, *tests*, *genetic*, *available*, *tested*, *how*, *performed*
INFORMATION	*information*, *about*, *provide*, *please*, *like*, *would*, *send*, *could*, *how*, *treatment*
MANAGEMENT	*treatment*, *treated*, *treatments*, *cure*, *treat*, *therapy*, *options*, *diet*, *effective*, *prevent*
MANIFESTATION	*symptoms*, *adpkd*, *among*, *dominant*, *polycystic*, *tend*, *autosomal*, *members*, *affected*, *kidney*
OTHEREFFECT	*between*, *problems*, *associated*, *syndrome*, *information*, *normal*, *issues*, *about*, *how*, *cause*
PERSONORG	*support*, *find*, *specialize*, *doctors*, *centers*, *contact*, *talk*, *area*, *groups*, *treating*
PROGNOSIS	*prognosis*, *life*, *expectancy*, *long*, *term*, *expect*, *affect*, *person*, *back*, *ability*
SUSCEPTIBILITY	*chances*, *genetic*, *inherited*, *rare*, *chance*, *information*, *child*, *passed*, *about*, *prevalence*

Table 3: Most associated words for each question type.

Fan Bu, Xingwei Zhu, Yu Hao, and Xiaoyan Zhu. 2010. Function-based question classification for general QA. In *Proceedings of EMNLP*.

Yllias Chali and Sadid A. Hasan. 2012. Simple or Complex? Classifying the Question by the Answer Complexity. In *COLING Workshop on Question Answering for Complex Domains*, pages 1–10.

Wen Chan, Weidong Yang, Jinhui Tang, Jintao Du, Xiangdong Zhou, and Wei Wang. 2013. Community question topic categorization via hierarchical kernelized classification. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1).

Huizhong Duan, Yunbo Cao, Chin-Yew Lin, and Yong Yu. 2008. Searching Questions by Identifying Question Topic and Question Focus. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 156–164.

John W. Ely, Jerome A. Osheroff, Mark H. Ebell, George R. Bergus, Barcey T. Levy, M. Lee Chambliss, and Eric R. Evans. 1999. Analysis of questions asked by family doctors regarding patient care. *BMJ*, 319(7206):358–361.

John W. Ely, Jerome A. Osheroff, Paul M. Gorman, Mark H. Ebell, M. Lee Chambliss, Eric A. Pifer, and P. Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *BMJ*, 321:429–432.

John W. Ely, Jerome A. Osheroff, M. Lee Chambliss, Mark H. Ebell, and Marcy E. Rosenbaum. 2005. Answering Physicians' Clinical Questions: Obstacles and Potential Solutions. *J Am Med Inform Assoc*, 12(2).

Ulf Hermjakob. 2001. Parsing and Question Classification for Question Answering. In *Proceedings of the ACL Workshop on Open-Domain Question Answering*.

Andrew Hickl, Kirk Roberts, Bryan Rink, Jeremy Bensley, Tobias Jungen, Ying Shi, and John Williams. 2007. Question Answering with LCC's CHAUCER-2 at TREC 2007. In *Proceedings of TREC-2007*.

Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2013. Interpreting Consumer Health Questions: The Role of Anaphora and Ellipsis. In *Proceedings of the 2013 BioNLP Workshop*, pages 54–62.

Tetsuya Kobayashi and Chi-Ren Shyu. 2006. Representing Clinical Questions by Semantic Type for Better Classification. In *Proceedings of the AMIA Annual Symposium*.

Vijay Krishnan, Sujatha Das, and Soumen Chakrabarti. 2005. Enhanced Answer Type Inference from Questions using Sequential Models. In *Proceedings of EMNLP*.

X. Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of COLING*.

Feifan Liu, Lamont D. Antieau, and Hong Yu. 2011. Toward automated consumer question answering: Automatically separating consumer questions from professional questions in the healthcare domain. *Journal of Biomedical Informatics*, 44(6).

Donald Metzler and W. Bruce Croft. 2005. Analysis of Statistical Question Classification for Fact-Based Questions. *Information Retrieval*, 8:481–504.

John M. Prager. 2006. Open-Domain Question Answering. *Foundations and Trends in Information Retrieval*, 1(2):91–231.

Kirk Roberts and Sanda M. Harabagiu. 2012. Locational Relativity and Domain Constraints in Spatial Questions. In *Proceedings of the ACM International Conference on Advances in Geographic Information Systems*.

Kirk Roberts and Andrew Hickl. 2008. Scaling Answer Type Detection to Large Hierarchies. In *Proceedings of LREC*.

Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. 2014. Annotating Question Decomposition on Complex Medical Questions. In *Proceedings of LREC*.

Connie Schardt, Martha B. Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med Inform Decis Mak*, 7(16).

Jessi Van Der Volgen, Bethany R. Harris, and Dina Demner-Fushman. 2013. Analysis of Consumer Health Questions for Development of Question-Answering Technology. In *Proceedings of the 2013 Annual Meeting and Exhibition of the Medical Library Association*.

Hong Yu and YongGang Cao. 2008. Automatically Extracting Information Needs from Ad Hoc Clinical Questions. In *Proceedings of the AMIA Annual Symposium*.

Hong Yu and Carl Sable. 2005. Being *Erlang Shen*: Identifying Answerable Questions. In *IJCAI Workshop on Knowledge and Reasoning for Answering Questions*.

Yan Zhang. 2010. Contextualizing Consumer Health Information Searching: An Analysis of Questions in a Social Q&A Community. In *Proceedings of the 1st ACM International Health Informatics Symposium*.

Improving the Extraction of Clinical Concepts from Clinical Records

Xiao Fu and Sophia Ananiadou
National Centre for Text Mining
School of Computer Science, University of Manchester
Manchester Institute of Biotechnology,
131 Princess Street, M1 7DN, Manchester, UK
E-mail: fux@cs.man.ac.uk, sophia.ananiadou@manchester.ac.uk

Abstract

Essential information relevant to medical problems, tests, and treatments is often expressed in patient clinical records with natural language, making their processing a daunting task for automated systems. One of the steps towards alleviating this problem is concept extraction. In this work, we proposed a machine learning-based named entity recognition system to extract clinical concepts from patient discharge summaries and progress notes without the need for any external knowledge resources. Three pre- and post-processing methods were investigated, i.e. truecasing, abbreviation disambiguation, and distributional thesaurus lookup, the individual annotation results of which were combined into a final annotation set using two refinement schemes. While truecasing and abbreviation disambiguation capture the inflectional morphology of words, the distributional thesaurus lookup allows for statistics-based similarity matching. We achieved a maximum F-score of 0.7586 and 0.8444 for exact and inexact matching, respectively. Our results show that truecasing and annotation combination are the enhancements which best increase the system performance, whereas abbreviation disambiguation and distributional thesaurus lookup bring about insignificant improvements.

Keywords: concept extraction, clinical records, truecasing, abbreviation disambiguation, distributional thesaurus

1. Introduction

The use of electronic medical records (EMRs) that contain clinical information on individual patients has been shown to reduce healthcare costs as well as improve the quality of healthcare provided (Holroyd-Leduc et al., 2011). It has become one of the most important novel technologies in the clinical domain (Murphy et al., 2007). However, clinicians and healthcare providers often find difficulties in filtering and retrieving useful knowledge from those clinical documents since the majority of EMR data is still expressed in narrative form (Pedersen, 2006; Xu et al., 2012a; Byrd et al., 2013). Therefore, concept extraction which has been actively employed to unlock information from free-text content (Denny et al., 2010; Gonzalez et al., 2012; Xu et al., 2012b) could be used to address this problem and consequently, to improve clinical care.

In this paper, we investigated a machine learning (ML)-based system to identify clinically relevant entities from patient discharge summaries and progress notes, and assign semantic types (i.e., problem, test, and treatment) to them, as specified by the concept extraction task of the 2010 i2b2/VA challenge (Uzuner et al., 2011).

2. Related Work

Recently, several studies have been conducted to apply rule- and/or ML-based approaches on EMRs to assist scientists in the extraction of valuable information. deBruijn et al. (2011) developed a discriminative semi-Markov hidden Markov model (HMM) based on a wide range of features generated from both training texts and external knowledge sources to identify clinical concepts in discharge summaries and progress reports. They observed that projecting textual features onto a high-dimensional feature space as well as the utilisation of external sources for semantic and syntactic tagging are beneficial. Jiang et al. (2011) proposed a hybrid clinical entity extraction system combining an ML-based named entity recogniser with rule-based methods for post-processing. Two ML algorithms, conditional random fields (CRF) and support vector machines (SVM) were applied. Their results suggest that CRF outperformed SVM, and the semantic features derived from existing medical knowledge bases can enhance the performance of clinical named entity recognition (NER) significantly. Similarly, Patrick et al. (2010) developed a hybrid medication extraction model which was based on a cascaded approach, incorporating two ML classifiers (i.e., CRF and SVM) and several pattern matching rules. In order to effectively use the two ML methods, they manually annotated another 145 records to augment the training set since they had access to only 17 annotated records initially. The performance of their model is better than those of the rule-based approaches adopted for the same task.

Instead of developing new ML- or rule-based methods, some researchers have focussed on combining existing approaches. Kang et al. (2011) integrated six named entity recognisers and chunkers, including both ML- and thesaurus-based methods, to annotate clinical records. Majority voting (Penrose, 1946), a method that validates if majority of the systems have given identical results, was then applied to generate composite annotations. They conclude that a combined annotation system for clinical records performs substantially better than any of the individual systems.

In our study, on the other hand, we did not leverage any external medical ontologies, such as UMLS (Bodenreider

et al., 2004) or SNOMED-CT (Lee et al., 2014). A relevant work is performed by Yang (2010), in which a rule-based medication extraction system for textual patient reports without employing any external medical knowledge resources was constructed. Seven types of medication information including drug names, dosages, modes of administration, frequencies, durations, reasons and contexts were extracted. Based on his findings, the rule-based approach, built based on a small, annotated development corpus and without utilising any knowledge bases, obtained satisfactory performance in the concept extraction task. In contrast, this paper investigates the recognition of different concept types and in clinical notes using a ML-based concept extraction system.

3. Methods

We initially constructed a CRF-based NER system to tag clinical records, which was considered as the baseline in our experiments. Several pre- and post-processing methods were then explored to improve the annotation performance.

3.1 Dataset

This study is performed on patient discharge summaries and progress notes in the 2010 i2b2/VA challenge data set (Uzuner et al., 2011). Seventy-three (73) human annotated records were used for system training, while the test data set was comprised of 256 annotated records. For each record, there are two types of files. One is the report file, which has already been split into sentences, and the other is the annotation file, in which annotations are specified by means of line and word numbers that indicate text spans corresponding to concepts. Table 1 shows a sentence in report files and its corresponding annotations in annotation files. In this corpus, there are 16,779 entities which were assigned the problem label, 12,261 assigned the test label and 12,417 annotated as treatment.

Sentences (Report File)	Trauma series demonstrated no evidence of a pelvic fracture.
Concept Annotations (Annotation File)	c='trauma series' 44:0 44:1||t='test' c='a pelvic fracture' 44:6 44:8||t='problem'

Table 1: Examples from a pair of report and annotation files

3.2 Baseline

We selected NERsuite[1], which is a freely available NER tagger based on the CRFsuite implementation of CRFs (Okazaki, 2007), to generate the concept annotations. NERsuite consists of three modules, a tokeniser, a modified version of GENIA tagger, and a NER. The procedure is as follows: First, the sentence-split report files were processed by the tokeniser which was used to segment each sentence into tokens, and compute the position of each token in the sentence. The modified GENIA tagger was then applied to produce three token features, i.e., the part-of-speech (POS) tags, lemmas, and chunk tags. In order to train the NER model, the annotated records were converted into a Begin, Inside, Outside (BIO) format to obtain the correct named entity (NE) label for individual tokens. Consequently, we used a total of seven possible NE labels, i.e., 'B-/I- problem, test, treatment', and 'O'. The example sentence in Table 1 will be transformed to the token-level annotation shown in Table 2. A CRF model was then trained to assign the label to each token in the test corpus.

3.3 Truecasing

Clinicians are used to writing short and terse sentences with limited use of full sentence syntax in clinical records, examples of which include *'Percocet for pain as needed. Aspirin 81 mg daily.'* and *'Patient may shower, no baths. No driving for at least one month.'*. While capitalisation is typically used only to begin sentences, we have observed that several full words are also capitalised for the purpose of emphasis. Given these considerations, we considered the application of the truecasing method, generally used to restore the correct case of tokens in raw texts to consistently transform token expressions to their canonical forms (Lita et al., 2003; Pyysalo and Ananiadou, 2013). The Truecase Asciifier module in Argo[2] (Rak et al., 2012a; Rak et al., 2012b) was employed to generate tokens in their normalised case form. Truecasing was performed before tokenisation (i.e., on input sentences) as it takes into consideration the context surrounding any given token.

3.4 Abbreviation Disambiguation

Acronyms and abbreviations also appear widely in clinical records, some of which follow certain conventions whereas others are ambiguous (Carroll et al., 2012). For instance, *q.d.* is the standard acronym for *'quaque die'*, which means once a day. *MI*, having more than 80 possible full forms, can stand for *myocardial infarction*, *mitotic index*, and *myo-inositol*, while *CAD* may mean any of *coronary artery disease*, *computer-aided diagnosis*, and *caldesmon* depending on the context. Therefore, a disambiguation process is crucial to ensure the correct interpretation of the records (Uzuner et al., 2011). In our study, we examined the impact of abbreviation disambiguation (AD) on NER by transforming the acronyms and abbreviations or short forms in the report files to their suitable expanded full forms estimated by Acromine Disambiguator[3], a word sense disambiguation classifier trained on MEDLINE abstracts (Okazaki et al., 2010).

3.5 Distributional Similarity

After we automatically annotated the test corpus using the newly trained NERsuite model, a distributional thesaurus,

[1] http://nersuite.nlplab.org/

[2] http://argo.nactem.ac.uk/

[3] http://www.nactem.ac.uk/software/acromine_disambiguation/

NE Label	Beginning Position	Past-the-End Position	Token	Lemma	POS	Chunk
B-test	0	6	Trauma	Trauma	NN	B-NP
I-test	7	13	series	series	NN	I-NP
O	14	26	demonstrated	demonstrate	VBD	B-VP
O	27	29	no	no	DT	B-NP
O	30	38	evidence	evidence	NN	I-NP
O	39	41	of	of	IN	B-PP
B-problem	42	43	a	a	DT	B-NP
I-problem	44	50	pelvic	pelvic	JJ	I-NP
I-problem	51	59	fracture	fracture	NN	I-NP
O	60	61	.	.	.	O

Table 2: The transformed annotations of 'Trauma series demonstrated no evidence of a pelvic fracture.'

in which each word is associated with a list of other words according to their distributional similarity (DS) scores (Carroll et al., 2012), was constructed using the training documents. The thesaurus was then applied to reassign concept types to tokens in the initial results for improving recall, based on the intuition that similar words tend to have the same concept type. In Figure 1 is a list of six words which were identified as most similar to 'artery' in the thesaurus. The numbers in the second column are DS scores that indicate the degree of similarity.

artery	artery	1.0
	embolism	0.1701
	insufficiency	0.1585
	angioplasty	0.1496
	coronary	0.1493
	percutaneous	0.1033
	⋮	

Figure 1: Distributional thesaurus entries for 'artery'

Pair-wise DS scores between words in the documents were computed using Lin's measurement (Lin, 1998), as shown in Equation 1.

$$sim(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\cap T(w_2)} (I(w_1,r,w) + I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w) + \sum_{(r,w)\in T(w_2)} I(w_2,r,w))} \quad (1)$$

where w and r represent words and the relationship between two words, respectively. $I(w,r,w')$ is equal to the mutual information between w and w' (see Equation 2).

$$I(w,r,w') = \log \frac{\| w,r,w' \| \times \| *,r,* \|}{\| w,r,* \| \times \| *,r,w' \|} \quad (2)$$

where $\| w,r,w' \|$ denotes the frequency of the triple (w,r,w') in the corpus, and $*$ means there are no specific words. $T(w)$ means the set of pairs (r,w') such that $I(w,r,w')$ is positive. Here we used the 4 types of proximity relationships as in the work of Carroll et al. (2012), shown in Table 3.

Relation Name	Explanation
prev	Previous word
prev_window	Word within a distance of 2-5 words to the left
next	Next word
next_window	Word within a distance of 2-5 words to the right

Table 3: Proximity relationships used to calculate DS

3.6 Hybrid Methods

Since a hybrid annotation system has been demonstrated to have a better performance than any of the individual systems (Kang et al., 2010; Uzuner et al., 2011), we combined the three aforementioned techniques (i.e., truecasing, AD, and DS) into a final annotation system by two schemes . Each of the schemes is illustrated in Figures 2 and 3, respectively.

The first is a sequential scheme in which the training and test documents were processed by the truecasing and the AD modules consecutively. In this way, we prevent the AD module from treating words in all uppercase letters as short forms and incorrectly expanding them. For example, AD will typically expand *'END'* to *'endurance'*, but with the application of the truecasing module, the former will be first transformed to *'end'*, hence avoiding its unnecessary expansion. Next, the texts with the correct case information and expanded forms were used to train and test the NERsuite model to obtain the concept extraction results. A distributional thesaurus built on the new training documents was then employed to assign new annotations to the test documents.

In the second scheme, a parallel one, we simply compute the union of the annotation results of these three systems. If the concept annotations provided by any two of the three systems are identical, they are generated as final, combined annotations. Otherwise, the result of truecasing is considered as the final annotation, since this technique performed best on the training texts.

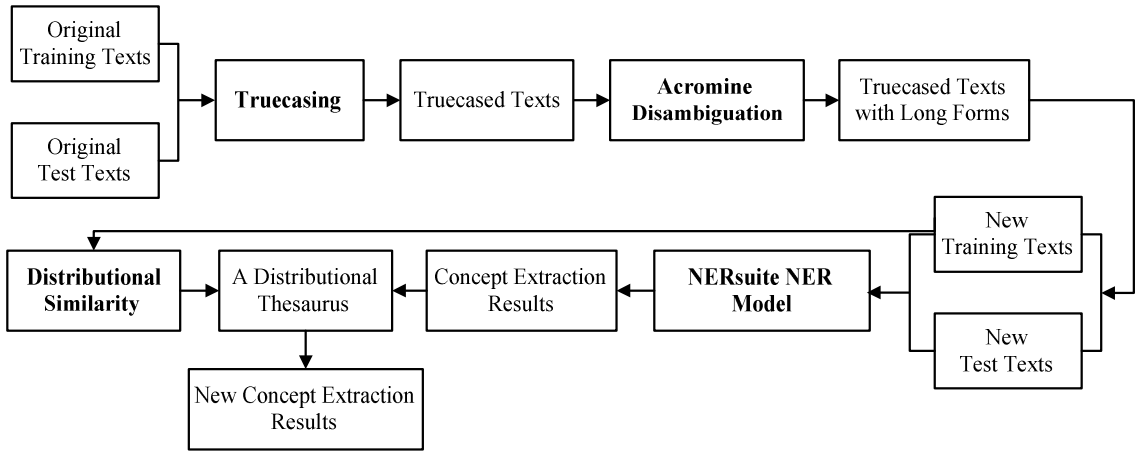


Figure 2: The framework of the sequential hybrid annotation system

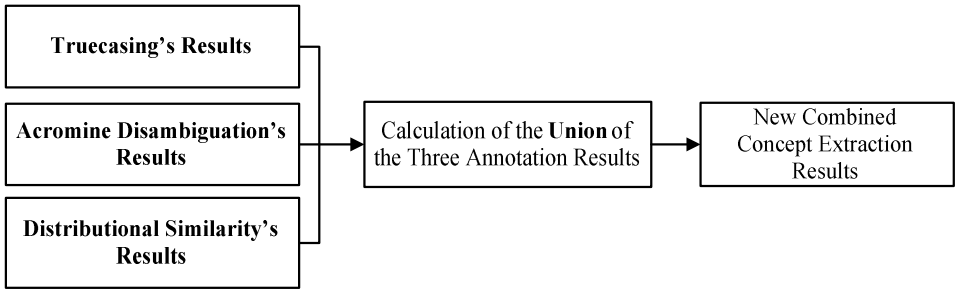


Figure 3: The framework of the parallel hybrid annotation system

4. Results and Discussion

4.1 Results

We explored the performance of truecasing, AD and DS alone, and various combinations of these three methods for clinical concept annotation. Table 4 summarises the performance of our systems on the 256 test records based on exact and inexact matching. The former requires that automatically generated annotations are exactly the same as those in the gold standard, while the latter additionally takes into account annotations with inexact boundaries as long as they have overlapping spans (Kang et al., 2010; Uzuner et al., 2011).

The truecasing method obtained an F-score of 0.7586 according to exact matching, making it our best performing concept annotation system. In contrast, the AD method did not show any improvements. While the DS method produced optimal recall (0.7225), precision was compromised (0.6466), resulting in a reduction in F-score. The F-score of the sequential combination of the truecasing and AD methods is 0.7556 which is worse than that of the purely truecasing-based system and even that of the baseline. The addition of the DS method further reduced the F-score to 0.6877. Nevertheless, integrating those three methods using the parallel scheme achieved better performance than the baseline and the parallel combination of truecasing and AD.

The best F-score measurements using inexact matching is 0.8444, achieved by the parallel combination of truecasing and AD. The parallel combination of truecasing, AD, and DS, truecasing on its own, and the sequential combination of truecasing and AD models outperformed the baseline as well.

Systems	Recall	Precision	F-score
Exact Matching			
Baseline	0.7132	0.8054	0.7565
T	**0.7147**	**0.8083**	**0.7586**
AD	0.7099	0.8014	0.7529
DS	0.7225	0.6466	0.6825
T + AD (S)	0.7115	0.8055	0.7556
T + AD + DS (S)	0.7210	0.6573	0.6877
T + AD (P)	0.7283	0.7849	0.7556
T + AD + DS (P)	0.8047	0.7140	0.7567
Inexact Matching			
Baseline	0.7927	0.8951	0.8408
T	0.7931	0.8970	0.8419
AD	0.7925	0.8947	0.8405
DS	0.8133	0.7280	0.7683
T + AD (S)	0.7922	0.8969	0.8413
T + AD + DS (S)	0.8138	0.7420	0.7762
T + AD (P)	**0.8140**	**0.8772**	**0.8444**
T + AD + DS (P)	0.7944	0.8953	0.8419

T, truecasing; S, the sequential scheme; P, the parallel scheme;

Table 4: The performance for clinical concept extraction

Concept Types	Problem			Treatment			Test		
	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
Exact Matching									
Baseline	0.7094	0.7757	0.7411	0.6846	0.8237	0.7477	0.7473	0.8292	0.7861
T	0.7138	0.7803	0.7456	0.6890	0.8271	0.7517	0.7419	0.8298	0.7834
AD	0.7071	0.7742	0.7392	0.6854	0.8196	0.7465	0.7385	0.8213	0.7780
DS	0.6562	0.7175	0.6855	0.6304	0.7527	0.6862	0.5110	0.5685	0.5382
T + AD (S)	0.7102	0.7793	0.7432	0.6881	0.8232	0.7496	0.7368	0.8252	0.7785
T + AD + DS (S)	0.7181	0.5596	0.6290	0.6974	0.7827	0.7376	0.7487	0.7124	0.7301
T + AD (P)	0.7245	0.7559	0.7399	0.7032	0.8063	0.7512	0.7589	0.8051	0.7813
T + AD + DS (P)	0.7099	0.7757	0.7413	0.6849	0.8251	0.7485	0.7493	0.8262	0.7859
Inexact Matching									
Baseline	0.8202	0.8970	0.8569	0.7450	0.8963	0.8137	0.8035	0.8915	0.8452
T	0.8237	0.9005	0.8604	0.7470	0.8967	0.8150	0.7982	0.8927	0.8428
AD	0.8188	0.8966	0.8559	0.7459	0.8919	0.8124	0.8039	0.8947	0.8469
DS	0.7851	0.8584	0.8201	0.8169	0.8560	0.7803	0.6178	0.6873	0.6507
T + AD (S)	0.8185	0.8982	0.8565	0.7471	0.8937	0.8139	0.8018	0.8980	0.8472
T + AD + DS (S)	0.8444	0.6579	0.7396	0.7626	0.8559	0.8065	0.8241	0.7841	0.8036
T + AD (P)	0.8412	0.8776	0.8590	0.7656	0.8779	0.8179	0.8259	0.8761	0.8503
T + AD + DS (P)	0.8219	0.8981	0.8583	0.7452	0.8977	0.8144	0.8066	0.8894	0.8460

Table 5: Concept extraction results for each concept type

Systems	Exact Matching				Inexact Matching			
	FN	(%)	FP	(%)	FN	(%)	FP	(%)
Baseline	711	2.14	379	1.14	579	1.75	298	0.90
T	706	2.13	372	1.12	586	1.77	294	0.89
AD	724	2.18	472	1.42	565	1.70	372	1.12
DS	647	1.95	1438	4.34	513	1.55	1270	3.83
T + AD (S)	707	2.13	477	1.44	558	1.68	377	1.14
T + AD + DS (S)	650	1.96	1111	3.35	501	1.51	909	2.74
T + AD (U)	621	1.87	541	1.63	464	1.40	443	1.34
T + AD + DS (U)	693	2.09	406	1.22	559	1.69	317	0.96

%: The number of unrecognised abbreviations/The total number of abbreviations in the test set*100

Table 6: The number of unrecognised abbreviations

We also assessed the extraction performance for each entity type (i.e., problem, treatment, and test), as shown in Table 5. Results for the test type indicate the highest F-scores, with around 0.77 using exact matching for all the systems except DS. Based on inexact matching, the systems excluding the sequential combination of truecasing, AD and DS achieved the best results in problem extraction with F-scores over 0.82.

In our system-level evaluation, truecasing obtained the best performance in recognising problems and treatments on the test records using exact matching, while none of the methods improved extraction performance for tests.

The highest F-score using inexact matching was also achieved by truecasing; adding AD to it employing the parallel combination scheme resulted to the highest improvements in both treatment and test extraction.

4.2 Discussion

In our study, several NLP methods were investigated to construct NER systems for clinical concept extraction. Instead of using large-dimensional bags of complex features and rules derived from the text itself and external sources (deBruijn et al., 2011; Jiang et al., 2011), we used only lemmas, POS tags and chunk tags generated by the modified GENIA tagger. Our approach offers the possibility to construct effective clinical concept annotation systems on a simple feature set, without using dictionaries or ontologies.

Recent studies have shown that the performance of combined or hybrid annotation systems is better than that of any individual systems (Kang et al., 2011). However, our experiment results are not able to support their findings. Only the parallel combination of truecasing and AD outperformed truecasing and AD individually; the performance of the other hybrid NER models fell in between that of the best and worst performing individual methods. This can be attributed to the possibly conflicting contributions of those three pre- or post-processing methods.

The truecasing method improved the concept extraction performance based on both exact and inexact matching,

demonstrating that correct case information is beneficial for entity recognition in clinical records. The expansion and disambiguation of abbreviations were supposed to decrease the size of acronyms and abbreviations in the set of false negatives. Unexpectedly, the false negatives generated by AD are even slightly greater than that from truecasing and the baseline methods (see Table 6). The low performance of AD can partly be explained by the fact that the Acromine Disambiguation tool which we used was trained on MEDLINE abstracts (Okazaki et al., 2010). Training on a suitable abbreviation disambiguation dictionary specifically geared towards terms in clinical records would likely enhance the performance of Acromine Disambiguation.

The DS method did not add much value to the performance of the NER models, and in certain cases, even reduced their performance significantly. For example, from Table 4 we found an 8.99 and 7.74 percentage points drop in F-score based on exact and inexact matching, respectively, when we added DS to the sequential combination of truecasing and AD. A possible explanation is that the amount of the training records used to build the distributional thesaurus is too small to cover the full breadth of term occurrence, which limited the thesaurus' efficacy. Thus, more annotated records need to be involved to build a high-quality distributional thesaurus.

5. Conclusion

In this study, we developed an ML-based model for clinical entity recognition and systematically evaluated the effects of three pre- and post-processing methods (i.e., truecasing, AD, and DS) which were combined using two schemes (i.e., sequential and parallel). Based on our results, the original model with the addition of truecasing achieved the best performance using exact matching with an F-score of 0.7586. Using inexact matching, the maximum F-score of 0.8444 was obtained by the parallel combination of the truecasing and AD methods.

The utilisation of a more sizeable clinical record data set for training the models can potentially improve the performance of the system. We plan to continue with the development of our system upon gaining access to such data sets; the MIMIC II database that contains clinical records for 32,077 patients (Saeed et al., 2002; Scott et al., 2013) could be considered as an alternative option.

6. References

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1), pp. D267--D270.

Byrd, R.J., Steinhubl, S.R., Sun J., Ebadollahi, S., and Stewart, W.F. (2013). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *International Journal of Medical Informatics*.

Carroll, J., Koeling, R., and Puri, S. (2012). Lexical acquisition for clinical text mining using distributional similarity. In *Computational Linguistics and Intelligent Text Processing*, pp. 232--246.

deBruijn, B., Cherry, C., Kiritchenko, S., Martin, J., and Zhu, X. (2011). Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5), pp. 557--562.

Denny, J.C., Peterson, J.F., Choma, N.N., Xu, H., Miller, R.A., Bastarache, L., and Peterson, N.B. (2010). Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *Journal of the American Medical Informatics Association*, 17(4), pp. 383--388.

Holroyd-Leduc, J.M., Lorenzetti, D., Straus, S.E., Sykes, L., and Quan, H. (2011). The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence. *Journal of the American Medical Informatics Association*, 18(6), pp. 732--737.

Jiang, M., Chen, Y., Liu, M., Rosenbloom, S.T., Mani, S., Denny, J.C., and Xu, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5), pp. 601--606.

Jonnalagadda, S., Cohen, T., Wu, S., and Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1), pp. 129--140.

Kang, N., Barendse, R.J., Afzal, Z., Singh, B., Schuemie, M.J., van Mulligen, E.M., and Kors, J A. (2010). Erasmus MC approaches to the i2b2 Challenge. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*.

Lee, D., de Keizer, N., Lau, F., and Cornet, R. (2014). Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association*, 21, pp. e11--e19.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, 2, pp. 768--774.

Lita, L.V., Ittycheriah, A., Roukos, S., and Kambhatla, N. (2003). tRuEcasIng. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1, pp. 152--159.

Murphy, E.C., Ferris, F.L., and O'Donnell, W.R. (2007). An electronic medical records system for clinical research and the EMR–EDC interface. *Investigative Ophthalmology & Visual Science*, 48(10), pp. 4383--4389.

Okazaki, N. (2007). CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite.

Okazaki, N., Ananiadou, S., and Tsujii, J. (2010). Building a high quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9), pp. 1246--1253.

Patrick, J., and Li, M. (2010). High accuracy information

extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5), pp. 524--527.

Penrose, L.S. (1946). The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1), pp. 53--57.

Pyysalo, S., and Ananiadou, S. (2013). Anatomical entity mention recognition at literature scale. *Bioinformatics*, btt580.

Pedersen, T. (2006). Determining smoker status using supervised and unsupervised learning with lexical features. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

Rak, R., Kolluru, B., and Ananiadou, S. (2012a). Building trainable taggers in a web-based, UIMA-supported NLP workbench. In *Proceedings of the ACL 2012 System Demonstration*, pp. 121--126.

Rak, R., Rowley, A., Black, W., and Ananiadou, S. (2012b). Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database: the journal of biological databases and curation*.

Scott, D.J., Lee, J., Silva, I., Park, S., Moody, G.B., Celi, L.A., and Mark, R.G. (2013). Accessing the public MIMIC-II intensive care relational database for clinical research. *BMC medical informatics and decision making*, 13(1), pp. 9.

Saeed, M., Lieu, C., Raber, G., and Mark, R.G. (2002). MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in Cardiology*, pp. 641--644.

Uzuner, Ö., South, B.R., Shen, S., and DuVall, S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5), pp. 552--556.

Xu, Y., Tsujii, J., and Chang, E. (2012a). Named entity recognition of follow-up and time information in 20 000 radiology reports. *Journal of the American Medical Informatics Association*, 19(5), pp. 792--799.

Xu, Y., Hong, K., Tsujii, J., and Chang, C. (2012b). Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5), pp. 824--832.

Yang, H. (2010). Automatic extraction of medication information from medical discharge summaries. *Journal of the American Medical Informatics Association*, 17(5), pp. 545--548.

Extracting Medical Concepts from Medical Social Media with Clinical NLP Tools: A Qualitative Study

Kerstin Denecke
University of Leipzig
Semmelweisstr. 14, Leipzig, Germany
kerstin.denecke@iccas.de

Abstract

Medical social-media provides a rich source of information on diagnoses, treatment and experiences. For its automatic analysis, tools need to be available that are able to process this particular data. Since content and language of medical social-media differs from those of general social media and of clinical document, additional methods are necessary in particular to identify medical concepts and relations among them. In this paper, we analyse the quality of two existing tools for extracting clinical terms from natural language that were originally developed for processing clinical documents (cTakes, MetaMap) by applying them on a real-world set of medical blog postings. The results show that medical concepts that are explicitly mentioned in texts can reliably be extracted by those tools also from medical social-media data, but the extraction misses relevant information captured in paraphrase or formulated in common language.

Keywords: Information Extraction, Natural Language Processing, Medical Social Media, Sublanguage Analysis

1. Introduction

The advances in internet and mobile technologies changed the way how people access, use and share information. Data and experiences are exchanged via instant messaging, blogs, social networking (e.g. Facebook) or video sharing (e.g. YouTube). These tools opened new ways of communicating and enabled for timeless and location-independent information exchange. Mayo Clinic researchers have opined that social media has begun a process of "revolutionising healthcare" by improving healthcare and quality of life (Aase et al., 2012).

In order to make use of the knowledge captured in this new information source, tools for automatic processing are necessary. Natural language as used in clinical documents or medical social-media is unstructured. In contrast, structured or normalised data is required for automatically analysing the content and to enable further interpretation and processing of the data. Extracting concepts (such as drugs, symptoms, and diagnoses) from clinical narratives constitutes a basic enabling technology to unlock the knowledge within texts and support more advanced reasoning applications such as diagnosis explanation, disease progression modelling, and intelligent analysis of the effectiveness of treatment (Jonnalagadda et al., 2012).

Algorithms and tools are already available for mapping clinical and biomedical documents to concepts of medical terminologies and ontologies (e.g. MedLee, MetaMap (Aronson, 2001), cTakes (Savova et al., 2009)). Once applied to a document they provide for extracted terms concepts of clinical terminologies that can be used to describe the content of a document in a standardised way. Existing tools for concept extraction were designed specifically to process clinical documents, i.e. they are specialised to the linguistic characteristics of these documents. The linguistic characteristics of clinical and biomedical texts have been analysed in painstaking detail by other researchers (Friedman et al., 2002), (Kovic et al., 2008), (Meystre et al., 2008). In contrast, the literary composition of medical social-media data has unfortunately not yet been analysed with the same degree of precision. Clinical texts such as radiology reports contain short, telegraphic phrases resulting in a compact description of facts and observations written in medical terminology. In contrast, medical social media texts can consist of complete, complex sentences (e.g. in blogs and forums) or can be very short without using complete sentences (e.g. Twitter). Consumer health vocabulary, clinical terms and common language is exploited to deliver information. Language in medical social-media differs from language in clinical documents. Table 1. summarises the linguistic characteristics of the two text types.

It is still unclear whether the clinical NLP tools are suited to process medical social-media data given the different language characteristics. This question will be addressed in this paper. We will assess the extraction quality of such tools through a qualitative study. The quality of two named entity recognition tools originally designed for processing clinical texts is compared when they are applied to medical social-media text.

The paper is structured as follows. Section 2. provides an overview on natural language processing in general and concept extraction in particular from clinical documents. Further, corresponding methods and tools will be reviewed. Then, we describe the data set and analysis method that was applied to study the extraction quality for the two tools MetaMap and cTakes that are applied to medical social media (section 3.). Section 4. describes the results of the assessment. In section 5., the results are discussed and observations on the content and linguistic characteristics of medical social-media are summarised. The paper finishes with conclusions and remarks on future work.

2. Extracting Information from Texts

In this section, methods and tools from extracting information from unstructured documents in general and from clinical documents in particular will be summarised.

Text type	Clinical text	Medical Social-Media
Sentence structure	ungrammatical sentences; short, telegraphic phrases (e.g.*Aspirin or Fever*); often without verbs or other relational operators	rather long sentences
Word usage	word compounds (*high blood pressure*), formed ad hoc; modifiers are related to temporal information (e.g. sudden), evidential information (e.g. rule out, no evidence), severity information (mild, extensive), body location	adjectives; descriptive and narrative words
Spelling	misspellings; abbreviations, acronyms	abbreviations, misspellings
Language	mix of Latin and Greek roots with corresponding host language (German, English); domain-specific language	common language, rather than domain-specific language or clinical terminology; host language
Semantic categories of words	Procedures, Disorders, Anatomy, Concepts and Ideas	Living Beings, Disorders, Chemicals and Drugs, Concept and Ideas

Table 1: Linguistic characteristics of clinical texts, and medical social-media

2.1. Methods for Information Extraction and NER

Information extraction identifies facts or information in texts (Grishman, 1998). An information extraction system is often specialised on a specific domain (e.g. medicine) (Grisham, 2002) and composed of several modules, normally working in a pipeline fashion (Cunningham, 2002). It requires lexical resources, that provide background knowledge and associated terms as well as domain knowledge. In the medical domain, multiple standardised vocabularies and ontologies are available (e.g. UMLS[1], SNOMED CT[2]). This knowledge is exploited by extraction tools to identify meanings in sentences and to identify relevant text snippets given an extraction task. Further, knowledge for interpreting the data is necessary.

Information extraction comprises several tasks, including named entity recognition, coreference resolution, relation extraction and template filling. Named-entity recognition (NER) aims at identifying within a collection of text all of the instances of a name for a specific type of thing (Cohen and Hersh, 2005) and is focus of this work. Examples of named entity categories in the medical domain include diseases and illnesses, drugs. More general named entity categories are person names, organizations, or locations. Recognised medical entities can be mapped to concepts of a medical terminology such as UMLS or SNOMED CT to enable a normalised representation of extracted information.

A concept represents a single meaning. Due to the flexibility in language usage, the same meaning can be expressed in different ways, e.g. through a noun, its synonym, an abbreviation etc. Through mapping of terms to concepts of a terminology, texts can be represented semantically and become interpretable for computer algorithms. For example, the UMLS Metathesaurus is organised by concepts: each concept has specific attributes defining its meaning. It is linked to the corresponding concept names in the various source vocabularies.

Entities can be recognised in natural language text in two ways:

1. a simple lexicon lookup; and
2. extraction patterns that are either manually created or learnt from training corpora using supervised machine learning techniques.

Lexicon lookup approaches search for matches with words of a lexicon of named entities in a given text. Difficulties are found to arise, namely because there is no complete dictionary for most types of medical or biomedical entities. Therefore, the simple text-matching algorithms that are commonly used in other domains are not sufficient here. In extraction pattern-based approaches, patterns such as "[Title] [Person]" for the extraction of a person name (e.g. "Mr. Warren") are generated either by hand or by supervised machine learning techniques. Manual rule-based approaches can be very efficient, but unfortunately such systems require manual efforts to produce the rules that govern them. Machine learning techniques on the other hand that do not require costly human annotators do however require large training corpora to train their underlying models.

For named entity recognition from unstructured texts, several tools exist. Stanford NLP tools[3], Alchemy API[4], LingPipe[5] or OpenCalais[6] are some examples for NLP tools that can be exploited for extracting named entities from unstructured text. However, these systems were mainly designed for processing news articles and often specifically trained on news data sets. They support detection of entities referring to persons, organizations or locations and are not designed for extracting diagnoses, medical conditions or medical procedures.

For these purposes, specialised tools were developed. Existing information extraction systems designed for processing clinical documents or biomedical literature are based on (1) pattern matching techniques such as regular expressions

[1] http://www.nlm.nih.gov/research/umls/

[2] http://www.ihtsdo.org/snomed-ct/

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

[4] http://www.alchemyapi.com

[5] http://www.alias-i.com/lingpipe

[6] http://opencalais.com

(e.g. the REgenstrief eXtraction tool (Friedlin and McDonald, 2006)), (2) full or partial parsing (e.g. LifeCode system (Mamlin et al., 2003), (3) a combination of syntactic and semantic analysis (e.g. MedLEE (Friedman et al., 1994) and MetaMap (Aronson, 2001)). They were mainly developed to extract information from textual documents in the electronic health record, among others from chest radiography reports, radiology reports, echocardiogram reports and discharge summaries. Evaluations showed that the current natural language processing tools for clinical narratives are effectively enough for practical use (Friedman et al., 2013).

2.2. Clinical NER: cTakes and MetaMap

In the following, two tools that were developed to process clinical text or biomedical literature are described in more depth. These tools are freely available and are used in our qualitative study.

The Apache Clinical Text Analysis and Knowledge Extraction System (**cTAKES**, (Savova et al., 2009)) is an open-source natural language processing system for extracting information from documents. The algorithms were specifically trained to process clinical documents. Among others, the system provides a recognizer that identifies clinical named entities in text using a dictionary-lookup algorithm. Through lexicon-lookup, each named entity is mapped to a concept of a terminology, e.g. the Unified Medical Language System (UMLS). The recognition concentrates on concepts of semantic types: diseases, sign/symptoms, procedures, anatomy and drugs. cTAKES was built using the Apache UIMA Unstructured Information Management Architecture engineering framework and OpenNLP natural language processing toolkit. Its components are specifically trained for the clinical domain out of diverse manually annotated datasets, and create rich linguistic and semantic annotations that can be utilised by clinical decision support systems and clinical research. cTAKES has been used in a variety of use cases in the domain of biomedicine such as phenotype discovery, translational science, pharmacogenomics and pharmacogenetics. In evaluations with clinical notes, the algorithm achieved F-scores from 0.715 to 0.76 (Kipper-Schuler et al., 2008).

The **MetaMap System** (Aronson, 2001) is provided by the National Library of Medicine (NLM). The tool maps natural language text to concepts of the UMLS Metathesaurus. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques. Besides being applied for both IR and data-mining applications, MetaMap is one of the foundations of NLM's Medical Text Indexer (MTI) which is being used for both semi-automatic and fully automatic indexing of biomedical literature at NLM. MetaMap was originally developed to extract information from MEDLINE abstracts, but has been applied to clinical documents as well (e.g. for pathology reports (Schadow and McDonald, 2003), for respiratory findings (Chapman et al., 2004)). MetaMap follows a lexical approach and works in several steps. First, it parses a text into paragraphs, sentences, phrases, lexical elements and tokens. From the resulting phrases, a set of lexical variants is generated. Candidate concepts for the phrases are retrieved by lexicon lookup from the UMLS Metathesaurus and evaluated. The best candidates are organised into a final mapping in such a way as to best cover the text. Precision of MetaMap, which is the fraction of retrieved concepts that are relevant, was assessed for different text types already, namely for respiratory findings (Chapman et al., 2004), mailing lists (Stewart et al., 2012) and figure captions in radiology reports (Kahn and Rubin, 2009). The precision for these text types ranges between 56% and 89.7%.

Ironically, while for online news a comparison of several NER tools (e.g., Alchemy API, DBpedia Spotlight, OpenCalais etc.) has already been performed (Rizzo and Troncy (Rizzo and Troncy, 2011)), there are yet no such similar comparisons made of NER tools for medical social-media. As such, evaluation results of NER tools in the medical domain are only available for extraction from clinical or biomedical texts while not for medical social-media. On the other hand, reliable technologies for analysing the textual content of medical social-media are necessary. Thus, we will analyse the mapping quality for two example clinical NLP tools on medical social-media data.

3. Methods

In this section, we describe our study design of a qualitative comparison of the two clinical named entity recognition tools (cTAKES, MetaMap), when they are applied to medical social-media documents. The objective is to clarify whether the tools extract relevant information from social media correctly and to determine which information remains unconsidered. The results of the study are important for the development of social media processing tools, in particular to decide whether existing technology is sufficient or whether and which adaptations are necessary to achieve good analysis results.

3.1. Data Set

We applied the two tools to 1) ten texts drawn from "Health Day News"[7] and 2) ten blog postings from "WebMD"[8]. The Health Day news service provides daily health news for both consumers and medical professionals. Content is provided by professional writers. Our data collection concentrated on information provided for consumers. WebMD provides valuable health information, tools for managing health, and support to those who seek information. The blog postings are collected from physician blogs made available through WebMD website, where physicians are writing about topics related to health and medicine. The postings were collected and HTML code was removed from the postings.

3.2. Study Design

The mapping results produced by MetaMap and cTakes when applied to the data set were checked manually, sentence by sentence by two persons, a computer scientist specialised in medical informatics and a medical doctor. They assessed different parts of the data set, thus no annotator agreement could be determined. The assessment of the output of the tools concerned the correctness of the extraction,

[7] http://consumer.healthday.com/

[8] http://www.webmd.com

the relevance of the extracted concepts and limitations and possibilities of the extraction tools with respect to processing medical social-media data. More specifically, the annotators had to judge the

- presence of the detected named entity (present in the text or not),
- relevance of the detected named entity (relevant or irrelevant), and
- type of the detected named entity (correct or incorrect).

We identified words that are crucial for understanding the text or sentence which could not be identified by either one of the tools used. Correct and incorrect annotations were counted. Extracted concepts were labeled as *wrong* when they did not represent the actual meaning of the underlying term or even when the extraction was incomplete (e.g. when for the phrase *breast cancer* only the concept referring to *breast* is provided). Further, observations on reasons for errors were collected.

The objective of the assessment is to give insights into the possibilities and limitations of these tools when they are applied to medical social-media data. Unfortunately, the systems use different versions of the UMLS. MetaMap was run with UMLS 2013AB, while cTAKES is distributed with UMLS 2011AB. However, this is not supposed to critically influence the mapping quality in our study.

MetaMap processing was restricted to identify concepts of semantic types that are referring to medical conditions, procedures, medications or anatomy. This restriction was made to achieve comparability with the cTakes results. CTakes only determines concepts referring to these semantic types. More specifically, MetaMap processing was restricted to the semantic types: Therapeutic or Preventive Procedure, Sign or Symptom, Physiologic Function, Pharmacologic Substance, Laboratory or Test Result, Laboratory Procedure, Injury or Poisoning, Disease or Syndrome, Diagnostic Procedure, Daily or Recreational Activity, Clinical Drug, Body System , Body Substance, Body Space or Junction, Body Part, Organ, or Organ Component, Body Location or Region, Behavior, Anatomical Structure, Activity, Acquired Abnormality. Table 2 shows to the cTakes categories the corresponding MetaMap categories as considered in this work.

cTakes Category	**MetaMap Categories**
Diseases	Disease or Syndrome; Acquired Abnormality
Sign / Symptom	Sign or Symptom; Physiologic Function; Laboratory or Test Result; Injury or Poisoning
Procedures	Therapeutic or Preventive Procedure; Laboratory Procedure; Diagnostic Procedure
Anatomy	Body System; Body Substance; Body Space or Junction; Body Part, Organ, or Organ Component; Body Location or Region; Anatomical Structure
Drugs	Pharmacologic Substance; Clinical Drug

Table 2: cTakes categories and MetaMap correspondence

4. Results

This section describes the evaluation results and observations of the annotators. Precision values determined in the evaluation are listed in Table 3.

cTakes achieved an average precision of 94% for the data set. Annotations of type *DiseaseDisorder, SignSymptom, Drug and Procedure* are correct with around 93%; and *Anatomy* annotations correct with an precision of 98%.

Compared to the cTakes results, MetaMap's results are more often incomplete and wrong. Symptoms are recognized best with an precision of 75.1%, followed by concepts referring to *procedures* with 69% precision. The precision values are significantly lower than those of cTakes.

One problem of MetaMap is that certain phrases such as *breast cancer* are not mapped to a disease or finding. Only the location (e.g. *breast*) is annotated. The annotators were asked to consider such mappings as incorrect.

Table 4 shows the proportion of extracted concepts per category. The tools show clear differences. cTakes extracted in total 1399 concepts. Half of them are referring to diagnoses and signs and symptoms. In contrast, MetaMap extracted 1020 concepts from the same data set and the majority of concepts refer to procedures and medication.

Relevant terms that were not annotated by MetaMap with corresponding concepts are for example: *fever, pregnant, autism, inflammation, cancer, tumor, blood pressure*. The pronoun *his* is mapped to the concept *Histidine*. Further, the verbs *said* and *led* are mapped incorrectly.

Other errors occur in both systems. It could be recognised that named entities referring to job positions, journals, or organisations used in the texts led to wrong or rather misleading annotations in both tools. For example from the phrase *director of the virus hepatitis program* the phrase *virus hepatitis* is annotated as disease occurrence. The term *division* in phrase *a member of the faculty at the division of global health at the University of California* is annotated as *Procedure Mention*.

Anatomical concepts occur sometimes in common language expressions (e.g. *don't have to go hand in hand*). The term *hand* in this phrase is annotated with the category *anatomical site mention* which is correct in general, but in that particular phrases no anatomical concept is meant. Another observation is that the annotation normally concentrates on medical concepts and looses the context. For example, cTakes annotates the phrase *drop in estrogen levels* with a concept referring to *estrogen level*, but the information captured in the complete phrase that this level is dropping is lost by such annotation. Another example is the annotation of the phrase *lack of sleep* where the annotation misses *lack*. Given the fact that cTakes concentrates

on extraction specific medical concepts, it is not surprising that also qualitative judgements such as *It's a very effective treatment* remain unconsidered in the mapping.
Additionally, to the categories that were actually target of the evaluation, cTakes provides annotations of type *Roman Numeral Annotation*. It could be recognized that abbreviations, personal pronouns or measurement units are mapped to concepts of that annotation type (e.g. the abbreviation *CDC* or the personal pronoun *I*). Almost all mappings to that type are wrong. The pronoun *I* is always annotated as Roman Numeral Mention which is a false positive annotation. Interestingly, number expressions such as *300ml* were correctly extracted and annotated.

Category	MetaMap	cTakes
Disease	59,6%	92,9%
Sign, Symptom	75,2%	92,9%
Procedure	69,05%	93,7%
Anatomy	54,08%	98,1%
Drug	66,54%	93,8%

Table 3: Precision per category for both systems

Category	MetaMap	cTakes
Disease	17.5%	35.7%
Sign, Symptom	12.6%	27.1%
Procedure	28.8%	19.1%
Anatomy	19.2%	15.5%
Drug	21.9%	2.4%

Table 4: Precision per category for both systems

5. Discussion and Conclusions

In this section, we discuss the results of the evaluation and limitations of the study design.

5.1. Discussion of the Evaluation Results

In summary, both tools are able to extract concepts from medical social media when medical conditions or procedures are explicitly mentioned and described by nouns. In particular the cTakes mappings are already sufficient to get an overview on the medical content of the posting. For a deeper understanding some additional information need to be provided.
In particular, both tools often fail in mapping or produce wrong mappings for verbs, personal pronouns, adjectives and connecting words. Clearly, these terms or at least their meaning and the relationships they infer, are relevant for interpreting the content of a sentence and text. Since persons are describing their own personal experiences and observations in medical social-media data, the language they use inevitably includes to a large extent verbs that describe activities of persons and personal pronouns; consequently, it is crucial not to lose the meaning of these personal accounts from patients or healthcare professionals while engaging in automatic processing of blog or forum content. Whereas missing or wrong mappings are not necessarily an algorithmic problem, they might be a problem of the underlying knowledge resource. For example, there is no concept representing the verbs *warn, recommend, cause* or for the adjectives *horrible, miserable, or ineffective* in the UMLS, the language resource on which the tested tools based. This is due to the fact that the terminology has been developed to formalize clinical knowledge, and thus the meanings of verbs or adjectives that are commonly used in medical social-media are unfortunately not covered by this terminology.
One must take into consideration that authors of medical social-media content often have no medical training. As a result they often do not use the proper medical terms, but paraphrase these concepts instead. People frustrated with their medical conditions may use a metaphor to refer to their maladies. For example, a cancer patient wrote: "The beast is going to kill me." While the metaphor "beast" is not normally considered as synonym for cancer, this is what the patient used to refer to his illness.
Extraction quality and completeness depends clearly on the content. When the text is dealing with diseases and clearly mentioned symptoms, both tools provide appropriate annotations. However, medical social-media texts discuss sometimes also issues related to nutrition or wellness. In that case, the tools are not performing well. The data set comprised two text types: one personal blog and health news. The blog postings were rather dealing with personal experiences. However, we could not recognise any quality differences in the mapping. When medical concepts are explicitly mentioned using the proper medical terms the tools identify them properly. In the medical social media data, verbs are bearing information, but the tools are not identifying them due to missing background knowledge on their meaning. E.g. the verbs *infected, elevated, transmitted* would be relevant to be determined.
Ambiguity of terms led to errors in both system. For example, the phrase *hand in hand* led to annotations of an anatomical concepts. Avoiding such errors is difficult and requires consideration of the larger context. For both systems, phrases referring to person names, organisations, journals or job positions led to errors. For example, the tools labeled concepts referring to *disease* or *prevention* when processing the phrase *center of disease prevention and control*. However, a correct annotation would determine the complete phrase.

5.2. Discussion of the Study Design

The problem of the large number of concepts not extracted with MetaMap originates in the restriction to some selected semantic type. The concept *breast cancer* belongs to the category *Neoplastic Process* which was not selected in the MetaMap settings. The same holds true for other terms where UMLS concepts exist, but our restriction in the semantic types led to missing mappings. We assume, that the mapping quality increases when some additional categories are included. This problem is obviously due to the study design. However, the restriction to certain semantic types is important to reduce mapping errors. MetaMap would otherwise identify too many irrelevant and wrong concepts similar to the extraction of type *Roman Numeral Annotation* in cTakes where almost each mapping was wrong. No information was found which UMLS semantic types are un-

derlying the cTakes categories. Otherwise, exactly the same types would have been included into the MetaMap settings. MetaMap provides mapping candidates and sometimes provides multiple, ranked mappings. We considered in the evaluation the best mapping or the first one in the list when multiple mappings had the same rank. It might be, that the correct mapping was not the first one.

We did not compare the mappings of the two tools directly, i.e. it was not identified to what extent the tools provided the annotation to the same words. Such comparison would be interesting for future analysis. Additional annotations of cTakes were not analysed in depth (e.g. annotation type "Predicate" that annotates verbs).

Given the huge amount of different medical social media sources, the selection of a representative data set for a study as presented in this paper is difficult. A broader spectrum of authors and multiple sources should be considered in future, in particular when improving the tools to ensure generality.

5.3. Potential Extensions of Clinical NLP Tools

What we can see from patients' everyday usage of language to describe maladies is that the classical synonyms for medical terms that exist in biomedical ontologies may be wholly insufficient for the data considered here. Consideration of metaphors, paraphrases and other ways that the lay population refers to illnesses and diseases could be a substantial extension of these ontologies when applying such extraction tools that rely upon biomedical ontologies to mine medical social-media postings. Given that relevant meanings that conform to how patients' articulate their symptoms are readily provided in more common vocabularies such as WordNet or consumer health vocabularies, some of the possible ways of improving the quality of mapping tools when processing medical social-media data is to consider additional knowledge resources or in the alternative to exploit a more general terminology. Consumer health vocabularies (CHV) which link everyday words and phrases about health (e.g., heart attack) to technical terms or jargon used by healthcare professionals (e.g., myocardial infarction) (Zeng and Tse, 2006), (Zeng et al., 2007), might in fact serve as a template for improving mapping tools for use in medical social-media. In fact, the open source, collaborative consumer health vocabulary initiative tries to develop a CHV for consumer health applications which is intended to complement existing knowledge in the UMLS.

Interestingly enough, in addition to terminology extensions such as those found in the CHV that augment the nomenclature of the UMLS, other improvements can likewise be made to mapping tools that are used in the medical social-media setting: 1) By including general terminological resources such as WordNet, meanings of adjectives could be recognized and considered in the analysis. 2) Another possibility against wrong mappings of medical social-media postings is to enhance the underlying ontology, but this must be done cautiously as it is a very complicated process and could probably lead to problems in processing professional language. 3) A third possibility for an improved mapping or for improved named entity recognition is the extension of the mapping algorithm. Aronson et al. (Aronson et al., 2007) showed that it is possible to apply successfully an ensemble of classification systems originally developed to process medical literature on clinical reports. Such approaches need to be assessed in the future to develop a better suited mapping tool for medical social-media.

In fact, various mapping tools could be used together to achieve a more complete extraction and annotation. Our evaluation showed that tools that extract organization or person names could help in avoiding mapping errors. Further, Open Information Extraction techniques could help in identifying relevant relations as they are expressed by verbs in medical social-media. This extraction paradigm learns a general model of how relations are expressed based on unlexicalized features such as part-of-speech tags (e.g., the identification of a verb in the surrounding context) and domain-independent regular expressions (e.g., the presence of capitalization and punctuation). By making use of such extraction techniques, it would no longer be necessary to specify in advance the relevant terms or patterns found in social media. This approach may prove more practical given the fact that medical postings are fast-changing, thus making it simply impossible to continuously update the language of social media and their underlying lexical resources manually. Such an approach of open information extraction could help to identify relations expressed by verbs in medical social-media which is so far impossible to do using existing mapping tools. To avoid wrong mappings of personal pronouns or connecting words, negative lists could be exploited, i.e. lists that instruct the algorithms not to map the listed words at all.

In sum, there are a number of obstacles that automatic processing tools must overcome in order to make better use of the richness of data found in medical social-media postings. Nevertheless, some of the methods we have analyzed in this chapter augur well for getting closer to meeting such challenges head on. In the end, better data extraction methods for medical blog content insures a healthier patient population and a more efficient healthcare delivery system.

We learned from the assessment that medical social media data contains named entities to a large extent referring to person names or organisations. Recognising person and organisation names in advance could help in reducing errors in concept mapping to clinical concepts. The named entities referring to persons and organisations could be filtered out before processing.

6. Conclusions

In this paper, we applied two existing clinical NLP tools MetaMap and cTakes to medical social media texts and assessed the mapping quality. Surprisingly, the number of wrong mappings were very low for the cTakes system. However, not all information relevant for an automated analysis and interpretation is made available by the cTakes mappings. Regarding linguistic characteristics of medical social media we learned, that in those texts named entities referring to persons and organisations occur frequently and require additional processing which is so far not realized by clinical NLP tools. In future, we will combine existing clinical mapping tools with general named entity recognition tools and concentrate also on relation extraction among

concept mentions.

7. References

Aase, L., Goldman, D., Gould, M., Noseworthy, J., and Timimi, F. (2012). *Bringing the Social-media Revolution to Health Care*. Mayo Clinic Center for Social-media.

Aronson, A. R., Bodenreider, O., Demner-Fushman, D., Fung, K. W., Lee, V. K., Mork, J. G., Neveol, A., Peters, L., and Rogers, W. J. (2007). From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. In *Biological, translational, and clinical language processing*, pages 105–112, Prague, Czech Republic, June. Association for Computational Linguistics.

Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLs Metathesaurus: The MetaMap program. In *Proceedings of the AMIA 2001*.

Chapman, W., Fiszman, M., Dowling, J., Chapman, B., and Rindflesch, T. (2004). Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. In *Stud Health Technol Inform.*, pages 487–91.

Cohen, A. and Hersh, W. (2005). A survey of current work in biomedical text mining. *Brief Bioinform*, 6(1):57–71, January.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, 36:223–254.

Friedlin, J. and McDonald, C. (2006). A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA Annu Symp Proc.*, pages 269–73.

Friedman, C., Alderson, P., Austin, J., Cimino, J., and Johnson, S. (1994). A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc.*, 1(2):161174.

Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35:222–235.

Friedman, C., Rindflesch, T. C., and Corn, M. (2013). Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of Biomedical Informatics*, 46(5):765 – 773.

Grisham, R. (2002). Chapter 30: Information extraction. In *Mitkov R: The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Grishman, R. (1998). Information extraction and speech recognition. In *Proceedings of the Broadcast News Transcription and Understanding Workshop, Lansdowne, VA*.

Jonnalagadda, S., Cohen, T., Wu, S., and Gonzalez, G. (2012). Enhancing clinical concept extraction with distributional semantics. *Journal of Biomedical Informatics*, 45(1):129 – 140.

Kahn, C. and Rubin, D. (2009). Automated semantic indexing of figure captions to improve radiology image retrieval. *Journal of the American Medical Informatics Association*, 16:280–286.

Kipper-Schuler, K., Kaggal, V., Masanz, J., Ogren, P., and Savova, G. (2008). System evaluation on a named entity corpus from clinical notes. In *Language Resources and Evaluation Conference, LREC 2008*, pages 3001–7.

Kovic, I., Lulic, I., and Brumini, G. (2008). Examining the Medical Blogosphere: An Online Survey of Medical Bloggers. *Journal of Medical Internet Research*, 10(3).

Mamlin, B., Heinze, D., and McDonald, C. (2003). Automated extraction and normalization of findings from cancer-related free-text radiology reports. *AMIA Annu Symp Proc.*, pages 420–4.

Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook Med Informat*, page 12844.

Rizzo, G. and Troncy, R. (2011). NERD: A framework for evaluating named entity recognition tools in the Web of data. In *ISWC 2011, 10th International Semantic Web Conference, October 23-27, 2011, Bonn, Germany*, Bonn, ALLEMAGNE, 10.

Savova, G., Bethard, S., Styler, W., Martin, J., Palmer, M., Masanz, J., and Ward, W. (2009). Towards temporal relation discovery from the clinical narrative. In *AMIA Annual Symposium Proceedings*, volume 2009, page 568. American Medical Informatics Association, American Medical Informatics Association.

Schadow, G. and McDonald, C. J. (2003). Extracting structured information from free text pathology reports. *AMIA Annu Symp Proc.*, page 584588.

Stewart, S., von Maltzahn, M., and Raza Abidi, S. (2012). Comparing MetaMap to MGrep as a tool for mapping free text to formal medical lexions. In *Proceedings of the 1st International Workshop on Knowledge Extraction and Consolidation from Social-media in conjunction with the 11th International Semantic Web Conference (ISWC 2012), Boston, USA, November 12, 2012*, pages 63–77.

Zeng, Q. and Tse, T. (2006). Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc*, 13(1):2429.

Zeng, Q., Tse, T., Divita, G., and et al. (2007). Term identification methods for consumer health vocabulary development. *J Med Internet Res*, 9(1).

Preliminary Evaluation of Passage Retrieval in Biomedical Multilingual Question Answering

Mariana Neves[1], Konrad Herbst[1,2], Matthias Uflacker[1], Hasso Plattner[1]

[1]Hasso-Plattner-Institute at the University of Potsdam, Germany
[2] University of Heidelberg, Germany
marianalaraneves@gmail.com, k.herbst@stud.uni-heidelberg.de

Abstract

Question answering systems can support biologists and physicians when searching for answers in the scientific literature or the Web. Further, multilingual question answering systems provide more possibilities and flexibility to users, by allowing them to write questions and get answers in their native language, and the exploration of resources in other languages by means of machine translation. We present a prototype of a multilingual question answering system for the biomedical domain for English, German and Spanish. Preliminary experiments have been carried out for passage retrieval based on the multilingual parallel datasets of Medline titles released in the CLEF-ER challenge. Two parallel collections of 50 questions for English/German and English/Spanish have been created to support evaluation of the system. Results show that the task is not straightforward, that additional semantic resources are needed to support query expansion and that different strategies might be necessary for distinct languages.

Keywords: question answering, multilingual, biomedicine.

1. Introduction

Question answering systems (QA) are important tools in the biomedical domain to provide answers to questions arisen from scientists and physicians. For instance, biologists frequently look for answers in the scientific literature or in the Web to confirm results obtained in the laboratory, e.g., whether a certain disease could be associated to a particular genetic mutation. However, answers to such questions can be scattered over different scientific publications (abstracts and full text), biological databases and Web pages.

Despite the importance of QA systems for both fields, not much previous work has been carried out for the biomedical domain, when compared to the state-of-art solutions in the medical domain (Athenikos and Han, 2010). Currently, there seems to be only three QA systems available for the biomedical domain (Bauer and Berleant, 2012): EAGLi[1], AskHermes[2] and HONQA[3]. However, both AskHermes and HONQA have a focus on the medical domain and might not be suitable for biologists. Nevertheless, this scenario has been changing in the last couple of years thanks to new initiatives such as the QA4MRE (Morante et al., 2012) and BioASQ (Partalas et al., 2013) challenges which support improvements in biomedical question answering.

Multilingual question answering systems offer a variety of new possibilities to the user, also for the biological domain. Although most of the relevant scientific publications are available in the English language, there is a myriad of other resources in other languages that could be explored, such as non-English biomedical scientific literature (e.g., Scielo[4] for Spanish and Portuguese), as well as the whole Web. Further, a multilingual QA system allows non-native English speakers to pose questions and receive answers in their native language while the system queries English (or any other language) documents by means of machine translation.

Previous research is scarce in the biomedical multilingual question answering field and in natural language processing (NLP) in general. From the QA systems cited above, only HONQA allows questions to be posed in other languages than English, namely Italian and French, whose performance has been evaluated in (Olvera-Lobo and Gutirrez-Artacho, 2011). The biomedical research seems to be always one (or more) step behind state-of-art in NLP for other domains due to the complexity of the matter and the lack of suitable resources for system development and evaluation. For instance, QA systems need to rely in high performing tools for the many pre-processing steps, such as tokenization, part-of-speech tagging and parsing. However, previous research have proved that sometimes specific models might be necessary for some domains, such as biomedicine (McClosky et al., 2010). One important step towards improvements in biomedical NLP is the recent EU-funded Multilingual Annotation of Named Entities and Terminology Resources Acquisition (Mantra) project (cf. Section 2.1.) which has supported improvements in named-entity recognition of biomedical terms during the CLEF-ER challenge last year. Finally, improvements in machine translation for the biomedical have also been recently published (Jimeno Yepes et al., 2013).

We present a prototype of a multilingual question an-

[1] eagl.unige.ch/EAGLi/
[2] http://www.askhermes.org/
[3] http://www.hon.ch/QA/
[4] http://www.scielo.org/

swering system for the biomedical domain and perform a preliminary evaluation for English, German and Spanish based on the resources released during the CLEF-ER challenge (Rebholz-Schuhmann et al., 2013). The QA system was developed on the top of SAP HANA, an in-memory database which include built-in text analysis functionalities for eleven languages. Additionally, we have created and made available two parallel collections of 50 questions each for English/German and English/Spanish. As far as we know, there is no multilingual parallel collection of questions nor previous evaluation for multilingual passage retrieval in biomedical QA systems. So far, the most comprehensive passage retrieval evaluation for the biomedical domain has been performed during the TREC 2006 and 2007 Genomics tracks (Hersh and Voorhees, 2009), but whose questions and background collection seems not to be made available after the end of the challenge.

In the next section, we will present an overview of the CLEF-ER resources followed by a description of the multilingual collection of questions. Then, we describe the question answering system which is under development (cf. Section 3.) and present an evaluation over the created collection of questions (cf. Section 4.). Finally, a discussion on the results and future work are presented in Section 5..

2. Data

2.1. CLEF-ER resources

The CLEF-ER challenge took place in 2013 as part of the Mantra project[5] which aims to provide multilingual documents and terminologies for the biomedical domain. For the scope of this challenge, Medline and patent documents have been released in five languages: English, German, French, Spanish and Dutch. Mapping to English documents were provided for all documents in each of the languages other than English but no mapping seems to exist between documents from two other languages, e.g., between Spanish and German.

For the scope of this work, we have utilized the Medline documents collections, which in fact consists only of the title of the documents, as background collection for our QA prototype. We have restricted our work to Spanish and German, given the support of the SAP HANA database for these languages (cf. Section 3.) and the knowledge of the authors on them. Table 1 illustrates examples of corresponding Medline document titles in English, Spanish and German.

Organizers have also released a terminology containing synonyms for terms in the above languages, which has been compiled from three resources: Medical Subject Headings (MeSH), Systematized Nomenclature of Human and Veterinary Medicine (SNOMED-CT) and Medical Dictionary for Regulatory Activities (MedDRA). An example of a term and its corresponding synonyms in other languages is shown in Table 2.

Table 3 summarizes the total number of Medline documents and synonyms per language within the terminology. These values are based on our own measurements after successful database import, thus, there might be discrepancies to the figures provided by the CLEF-ER organizers[6]. The CLEF-ER terminology contains a total of 525,794 terms which can be associated to synonyms in the many of the supported languages. For instance, in Table 2, one concept is shown (C0000119) which contains 7 synonyms: 3 for English, one for French, one for German and two for Spanish.

Resource / Language	English	German	Spanish
Medline documents	1,593,546	719,232	247,655
Synonyms	1,771,498	118,902	622,638

Table 3: Number of Medline documents and synonyms for English, Spanish and German in the CLEF-ER resources.

2.2. Questions collection

The construction of the collection of parallel questions was based on the Medline documents in Spanish and German which contain a corresponding document in English. We chose this approach to allow a comparison between results obtained with German and Spanish with those in English on the same documents.

Batches of 50 documents (titles) were randomly retrieved from the above datasets and titles were manually chosen according to their relevance to the biomedical domain. During the automatic retrieval of candidate documents, we ignored titles which had length lower than 150 characters as they might not contain enough information for question generation. During manual screening of the titles, we avoided documents related only to the medical domain, an area where state-of-art in question answering is more advanced in comparison to the biological one (Athenikos and Han, 2010). In particular for the Spanish questions, we selected titles related to tropical neglected diseases, which is a frequent topic on publications in the Medline collection for this language. Finally, we ignored titles which were not relevant to the biomedical domain, such as "On the effect of religious schools on values of young people." (document d5582363), or that consisted of reports on a meeting or on the current situation in a particular place, such as "The 5th Annual Meeting of the Swiss Association for Preventive and Restorative Dentistry (SVPR) of 13 November 1999 in Zurich." (document d10744522).

Questions were manually written in a way that at least one answer could be found in the corresponding document, i.e., the document's title. However, not all of the information cited in the text was always used and the questions sometimes are more general than the respective text. We have generated only factoid questions, i.e., questions which

[5] https://sites.google.com/site/mantraeu/home

[6] https://sites.google.com/site/mantraeu/access-content

English/German (document d11951797)
Optimal intravascular brachytherapy: safety and radiation protection, reliability and precision guaranteed by guidelines, recommendations and regulatory requirements. Optimale intravaskuläre Brachytherapie. Sicherheit und Strahlenschutz, Zuverlässigkeit und Präzision gewährleistet durch Leitlinien, Empfehlungen und Verordnungen.
English/Spanish (document d18959013)
Impact of the deep breathing maneuver in the gas exchange in the subject with severe obesity and pulmonary arterial hypertension associated to Eisenmenger's syndrome. Impacto de la maniobra de inspiracin profunda en el intercambio gaseoso del sujeto con obesidad severa e hipertensin arterial pulmonar asociada a sndrome de Eisenmenger.

Table 1: Examples of Medline document titles from the CLEF-ER collection for the pairs English/German and English/Spanish.

[Term] id: C0000119 name: 11-Hydroxycorticosteroids namespace: mesh_term_from_umls def: "A group of corticosteroids bearing a hydroxy group at the 11-position." [] synonym: "11 Hydroxycorticosteroids" EXACT SYN_EN [MSH:D015062] synonym: "11-Hidroxicorticoesteroides" EXACT PREF_ES [MSHSPA:D015062] synonym: "11-Hidroxicorticosteroides" EXACT SYN_ES [MSHSPA:D015062] synonym: "11-Hydroxycorticosteroide" EXACT PREF_DE [MSHGER:D015062] synonym: "11-Hydroxycorticosteroids" EXACT PREF_EN [MSH:D015062, NDFRT:N0000011376] synonym: "11-Hydroxycorticosteroids [Chemical/Ingredient]" EXACT SYN_EN [NDFRT:N0000011376] synonym: "11-Hydroxycorticostrodes" EXACT PREF_FR [MSHFRE:D015062] is_a: C0020343 ! Hydroxycorticosteroids relationship: has_semantic_type T110 ! Steroid relationship: has_semantic_type T125 ! Hormone

Table 2: Term "1-Sarcosine-8-Isoleucine Angiotensin II" from the CLEF-ER terminology and its corresponding synonyms in English, French, Spanish and German.

requires one or more specific short answer in return, such as a chemical compound, an organism or a disease. While writing the questions, we tried to rephrase the text, used synonyms for both the named entities and remaining words (whenever possible), changed the word's lexical class (e.g., from verb to noun), and converted passive voice to active voice, or the other way round, following procedures described in (Heilman and Smith, 2010). Synonyms for the lexical terms were supported by making queries to a variety of on-line language-specific dictionaries.

The semantic concepts referred in the text were also, whenever possible, changed to a equivalent synonyms. This task was supported by a variety of on-line resources, such as Wikipedia (for the three languages), NCBI Taxonomy and other web sites which were returned by the Google search engine. Therefore, we did not make use of the thesaurus made available by the CLEF-ER challenge, instead, we have tried to use the resources that the users (biologists) might use while posing questions to a QA system.

Questions were initially written in English and were reviewed by an expert in molecular biotechnology (KH). In a second step, the questions were translated into Spanish and German by the authors, who are either native or have advanced knowledge on the languages. We have also sought to use synonyms for the words and concepts in this step. Finally, we tried to make the questions with similar difficulty level in both languages. Some examples of questions are presented in Table 4 and the list of parallel questions is available for download[7].

English/German (document d6357751)
Which methods can be used to determine the living cell count of cariogenic microorganisms? Welche Methoden bieten sich zur Bestimmung der Lebendzellzahl von kariogenen Mikroorganismen an?
English/Spanish (document d16888692)
What are possible drug targets for eye related infections? Cuáles son los posibles objetivos farmacológicos en infecciones relacionadas con el ojo?

Table 4: Examples of parallel questions from the English/German and the English/Spanish datasets.

[7] https://sites.google.com/site/marianalaraneves/resources/

3. System architecture

Question answering systems are usually composed of three steps (Athenikos and Han, 2010): question processing, passage retrieval and answer processing. In the first step, the system identifies the type of question which has been posed (e.g., yes/no, definitional or factoid), the expected answer e.g., gene/protein, disease, etc.), in case of factoid questions, and converts the question into a query to the passage retrieval step. It might also include identification of semantic concepts and query expansion based on available lexical thesaurus (e.g. WordNet) or domain-specific resources (e.g., UMLS). In the passage retrieval step, queries are posed to a collection of documents and passages, usually a couple of sentences, are ranked according to their relevancy to the query.

In this section, we describe the QA system prototype that is under development and that has been used for a preliminary evaluation of the parallel collection of questions. For this work, we focus on the question processing and passage retrieval steps, as we do still do not provide suggestions of answers for the proposed questions.

3.1. Question processing

Our prototype system starts with the tokenization of the question followed by the part-of-speech tagging of resulting tokens. We use the OpenNLP Maximum Entropy-based tokenizer and part-of-speech tagger and the corresponding available models for English and German[8]. Given that no models are available for the Spanish language, we have used the one available for Portuguese for the tokenization step, given the similarity between these languages and that tokenization is even more challenging in the later due to the composed words, which are not common in Spanish. For the part-of-speech tagging in Spanish, we have used the Maxent model made available by Juan Manuel Caicedo Carvajal[9].

Some tokens were filtered out according to language-specific Stopwords lists and to the part-of-speech tags which indicates numerals, e.g., "CD" for English, "CARD" for German, "DN" and "Z" for Spanish. For the later, we have used existing lists available for English[10], German and Spanish (both from the stop-words project[11]) and we have extended them occasionally with extra words which were missing, such as the prepositions "de" and "al" from the Spanish.

We carry out a query expansion of the tokens using the thesaurus made available in the CLEF-ER challenge (cf. Section 2.1.). The CLEF-ER thesaurus has been loaded into the HANA database and a fuzzy search is carried out for each token in the query against the synonyms available for the corresponding language. Comparison of the query term and the synonyms is performed by requiring at least 90% similarity of the terms, to allow more flexibility of the matching. For performance reasons and in order not to include potential irrelevant synonyms, we only expanded the query with the synonyms identified as preferred (e.g, "PREF_EN", "PREF_DE") in the CLEF-ER terminology (cf. Table 2).

Weights are assigned for the terms of the query and the corresponding synonyms and are calculated based on the popularity of the term in the CLEF-ER terminology. The higher the number of synonyms which match to a term, the lower the weight of the later. Tokens which do not match any synonyms are assigned weight 0.5, i.e., an average weight. Otherwise, weights are calculated based on the number of terms which matched to this particular token (#MatchesToken) and the total number of terms matched to all tokens of the query (#MatchesTotal), following the expression below:

$$weight = 1 - \frac{\#MatchesToken}{\#MatchesTotal} \quad (1)$$

3.2. Passage retrieval

Passage retrieval is performed using the SAP HANA database[12] (hereafter called HANA), an in-memory database which has already been successfully employed for real-time analysis of biomedical data (Schapranow et al., 2013). HANA provides built-in text analysis functionalities for eleven languages (Arabic, simplified Chinese, Dutch, English, Farsi, French, German, Italian, Japanese, Korean, Portuguese, Spanish) which includes integrated sentence splitting, tokenization and stemming components during full text indexing of the documents. Further, it allows fuzzy searching the full text index according to a pre-defined percentage (e.g., 90%) and also linguistic searching by considering linguistic variants of the query terms.

The Medline documents available for the English, Spanish and German languages from the CLEF-ER (cf. Section 2.1.) were loaded into HANA and a full text index was created for each collection. We have experimented with three search strategies provided by HANA: exact, fuzzy matching (at least 90% of similarity) and linguistic. When considering the linguistic search, which allows matching of terms which share the same stem, we apply it only to the terms of the query which have not been matched to any synonym. When querying for relevant passages, HANA proceeds in matching the terms to the individual tokens derived from the full text index of the documents according to the specified search strategy, whether exact, fuzzy or linguistic. A score is calculated by HANA for the matching tokens and sentences are ranked according to the weighted sum of scores of all matching tokens per sentence. As HANA automatically split the sentence during indexing of

[8] http://opennlp.sourceforge.net/models-1.5/

[9] http://cavorite.com/labs/nlp/opennlp-models-es/

[10] http://www.textfixer.com/resources/common-english-words.txt

[11] https://code.google.com/p/stop-words/

[12] http://www.saphana.com/

the documents, the current system could also be applied to abstracts of full text documents and not only to titles (usually single sentences). Table 5 shows the top ten sentences retrieved for the one of the English questions shown in Table 4.

4. Experiments and results

We randomly split the two parallel collections of 50 questions in English/Spanish and English/German in two sets of 25 question each, i.e., two sets of 25 questions for development and two sets of 25 questions for testing purposes. The development datasets were used for choosing the best strategies, adding of extra stopwords and error analysis, but have not been used to train any of components of the system.

Evaluation of the development and test datasets consisted in running the system for each of them, retrieving the 10 best ranked passages (sentences) and checking whether the original document from which the question had been derived was present in this list. With such an experiment, we sought to evaluate the precision of our QA prototype for finding relevant passages to the questions as well as the difficulty level of the questions. Text passages were retrieved only for documents in the same language of the question, e.g., Spanish questions were queried only against the Spanish documents, thus, no machine translation was used.

We have evaluated the following settings of our QA system:

1. HANA exact search;
2. HANA fuzzy matching (at least 90% similarity);
3. HANA fuzzy matching (above), plus query expansion of the question words using the CLEF-ER terminology;
4. HANA fuzzy matching (above), query expansion (above), plus HANA linguistic matching for those words in the question which did not match to any synonym.

For the evaluation, we calculate the R-Precision, which is the precision on the r-th position where a first match with the original document was found, or zero, if not found. For instance, if the first match is found in the third position, the R-P is 0.33 (1/3). We then calculate the mean of the R-precision over the collection of questions, i.e., over 25 questions for each dataset. Table 6 shows the results for each language in the English/German and English/Spanish parallel collections of questions.

Settings	English-German		English-Spanish	
	EN	DE	EN	ES
exact	0.04 (1)	0.04 (1)	0.01 (1)	0.09 (4)
fuzzy	0.05 (3)	0.02 (1)	0.10 (5)	0.10 (5)
+ synonyms	0.09 (3)	0.06 (2)	0.11 (4)	0.09 (6)
+ linguistic	0.10 (4)	0.04 (2)	0.03 (2)	0.08 (6)
Test	0.05 (5)	0.05 (4)	0.05 (3)	0.06 (4)

Table 6: R-Precision and number of sentences found (in parenthesis) for each dataset and for various setting of the question answering system. Results for the training dataset are shown for the many setting options while the ones for the test dataset were obtained when using query expansion and fuzzy matching (i.e., "+ synonyms"). Codes for the languages are the following: "EN": English, "DE": German, "ES": Spanish.

5. Discussion and future works

In this work, we have presented a preliminary evaluation of a passage retrieval component for a multilingual question answering system on two datasets of 50 parallel questions for English/German and English/Spanish. Regarding the construction of the question collection, we believe that 50 questions per collection (100 in total) is an adequate number for evaluation purposes, given that previous challenges in this field have utilized datasets with similar size, such as the 40 questions from the in the machine reading dataset for Alzheimer Disease (Morante et al., 2012) and the batches of 100 questions released during the BioAsq challenge (Partalas et al., 2013). As described in Section 2.2., the selection of the questions was carried using a hybrid approach of first randomly retrieving batches of 50 document titles from the CLEF-ER Medline collection and then manually choosing those which were more related to the biological domain and in order to avoid irrelevant titles. We define a relevant title as those that contain enough information to build an interesting factoid question without the need to refer to the abstract or the full paper of the publication. We believe that this approach ensures both the variability of our dataset, through the random selection of the candidates, as well as its quality through a subsequent manual selection of relevant titles.

Deriving questions from the original document titles required deciding which entity type, whether a species or a disease, was particularly interesting to be the subject of the answer. Writing the question collection was a task that took from 5 to 10 minutes per question, depending on the difficulty in rephrasing the text and in finding appropriate synonyms for the biological terms and remaining words, which required authors to refer to several on-line resources. For instance, for a certain question, "polio" has been used as synonym to "poliomyelitis" and "factor favoring" has been rephrased as "cause increased severity". Acronyms were frequently used whenever available, such as "HCV" instead of "hepatitis C virus". In some cases where no adequate synonym could be found for a term, we have opted for using hypernyms instead, such as "fish" in substitution for "tilapia", which requires question answering systems to consider hierarchical relationships between terms. Other types of relationships have been also explored, such as referring to a organ instead of a disease name, for instance, "eye related infections" instead of "keratoconjunctivitis".

Which methods can be used to determine the living cell count of cariogenic microorganisms?
1. Determination of the living cell count of cariogenic microorganisms using the measurement of their ATP content in the bioluminescence procedure–a critical look at the method. 2. Comparison of the coulter-counter-method and the counting-chamber to determine the cell count of the cerebrospinal fluid. 3. Phase contrast microscopic studies of living germs, a supplemental method for the study of effect of germicidal substances on microorganisms. 4. Germ count determination" from water samples by the plate method with special reference to the pH value of the used nutrient media. 5. Schistocytes: which definition should be taken and which method should be used to identify and count them?. 6. Comparative survey concerning methods of vestibular exploration used in 10 Western European countries. 7. The methods used to collect hematopoietic stem cells. 8. Comparative evaluation of the methods used to determine the sensitivity of bacteria to antibiotics. 9. Critical study of various methods used to determine the activity of anti-typhoid vaccine. 10. Human bone marrow cell culture–a sensitive method for the evaluation of the biocompatibility of materials used in orthopedics.

Table 5: Retrieved passages for a English question. The original title from which the question has been derived is the one in the first position of the rank.

In general, preliminary results show that passage retrieval is not a straightforward task for none of the three languages, even when using a small collection of Medline document titles. Although the best results were obtained for Spanish, this was probably due to the smaller set of documents available for this language (cf. Table 3) than to the simplicity of the task or the language. On the other hand, results for German were particularly low in comparison to the other languages because of the presence of many compound words, as will be discussed below. Despite the overall low performance of our results, these experiments provide baseline results for the proposed collection and gives an estimation on the quality of our dataset.

In comparison to the development dataset, performance for the test dataset was higher for the English/German dataset and a little bit lower for the English/Spanish questions. Curiously, this is the best performing result for the German questions. Discrepancies between development and test data are expected given the small number of questions and the distinct topics they refer. The development dataset was used only for comparing the different setting of the features, updating the stopwords list with additional terms and performing an analysis of the results.

Results vary significantly for the many configurations of our system when evaluated for the three languages (cf. Table 6). Despite the less complexity of the English language, more exact matches have been obtained for the Spanish (4) than for English (1 for each dataset) or German (1). The higher number of matches for Spanish have occurred mainly due the impossibility in getting better synonyms in the Spanish language when writing the question, together with the fact that less documents are available for this language. For instance, the question "Qué métodos histoquímicos permiten la observación y caracterización de glicoconjugados?" got the right match (document d9279022) in the third rank, while no correct match was found for the corresponding English question ("Which histochemical methods allow observation and characterisation of glycoconjugates?").

As expected, consideration of fuzzy matching (at least 90% of similarity) improves both the average R-Precision and the number of retrieved documents for all languages (except for German), as more terms can be matched using this approximate matching. Additionally, this improvement came with no degradation of the R-Precision, which also increased for all languages, again, with the exception of German. For the later, the only question which got a match in the first rank when using exact matching, "Spielt die Methode der RNA in situ Hybridisierung bei Studien an Rhesusaffen eine Rolle?" (document 7534003), this time got a match in the second rank, thus the decrease in precision. This was due to the fuzzy match between the tokens "HIV-1-RNA" and "RNA", as words are always tokenized by the dashes in HANA.

Contrary to the expected, the query expansion brought just one additional match for the German and Spanish datasets, but also one less for one of the English datasets, while not changing the other one. On the other hand, it increased the average R-Precision for most of the datasets, except for the Spanish one. This improvement occurred also because of the weights associated to each term, which help giving higher score to sentences containing terms with associated synonyms, as opposed to common words of the language not related to biomedical domain. As an example of an improvement due to query expansion, the short Spanish question "Qué patógenos de mosquitos existen?" ("Which mosquito pathogens exist?") got no matches in the fuzzy search but got one when using query expansion due to the addition of the synonym "Culicidae". The same synonym was retrieved for the corresponding English question but the large collection of documents make it more difficult to get the correct document in the top results for such a general question. However, many the documents which were returned to this English questions could also be

considered as correct, only that none of them was the one originally used for creating the question. Thus, future work could include the expansion of the questions datasets with other relevant passages (document titles) other than the original one.

Nevertheless, query expansion did not help much in getting new terms due to the limitations of the CLEF-ER terminology which is based only on three resources (cf. Section 2.1.) and without the exploration of the relationships between the terms, i.e., hypernyms and hyponyms. For instance, the adjective "corneal" was not obtained for the term "eye", although the synonym "Corneal Disease" does exist in the terminology. Another example of synonyms which could not be found only relying on the CLEF-ER thesaurus is the pair "endovenous" and "intravenous", none of them is present in the resource. Further, the use of a token-based query expansion did not allow the correcting matching of the terms, along with the complexity of the biomedical nomenclature. For instance, it is not straightforward to match "Staphylococcus aureus" to "S. aureus" without consideration of pattern rules specific for the species nomenclature.

Finally, we studied the use of HANA linguistic search for those terms which did not match synonyms in the CLEF-ER terminology, which we expected to be common words of the language instead and named-entities. However, this search approach did not get additional document matches due to the way that the score is calculated by HANA when using this type of search. The only additional match we got for an English document seems not to be directly related to this feature.

An analysis of the questions to which the corresponding documents could not be found in the top 10 ranking results gives us insights on the next steps to improve our system. First of all, different strategies might be necessary for different languages. While token-based search successfully obtained some matches for English and Spanish, it failed to perform well for German due to the high number of compound words. For instance, if a question contains the word "Pankreaskarzinom" and the document the term "Adenokarzinom" neither an exact nor 90% fuzzy matching would be able to get such match, while the word "carcinom" (in English) and "carcinoma" (in Spanish) would match in the corresponding questions and documents in these languages. For future work, we want to explore advanced natural language processing functionalities of the HANA database which identifies compound words for German. For instance, the word "hochdosiert" (high dosed), which is an adjective derived from the adjective "hoch" (high) and the verb "dosieren" (to dose), is identified in the HANA full text index as "hoch#dosieren", thus allowing partial matches to any of the words in its composition.

We have relied in external libraries (i.e., OpenNLP) for the pre-processing of the questions, which has been also limited to the models available for the corresponding languages. As future work, we will study the use of the HANA database for this step as well, given that it already provides support for these languages as well as shallow parsing information (chunks). The later has not been explored in this work but can certainly help in both the query expansion, when matching query terms to potential synonyms, as well as in the passage retrieval step, when matching terms to documents. Further, other metrics for the terms weights and a range of values for the fuzzy matching will also be studied.

Regarding the semantic resources, future works will certainly use additional multilingual terminologies, such as DBpedia[13], as well as the exploration of semantic relationships between the terms for obtaining also hypernyms and hyponyms. Also additional lexical resources will be explored for retrieving synonyms and related words to common words of the languages, such as Wordnet for English, and similar resources for the other languages. In this work, the background collection has been limited to the document titles made available during the CLEF-ER challenge, given the scarce resources of biomedical documents in languages other than English. Therefore, future work could also explore the use of machine translation (Jimeno Yepes et al., 2013) for translating question in other languages to English, thus being able to query the English documents in Medline.

In this work, we have limited languages to English, German and Spanish according to the languages which the authors were more confident, as well as the ones supported by both the CLEF-ER resources and the HANA database. However, changes in the systems for any of the other languages already supported by HANA (Arabic, simplified Chinese, Dutch, Farsi, French, Italian, Japanese, Korean, Portuguese) would only require a background collection of documents and a pertinent list of questions.

6. Conclusions

In this work we have presented our prototype of a multilingual question answering system for the biomedical domain and have evaluated the system using two collections of 50 questions for English/German and English/Spanish. Our system works on the top of a SAP HANA database which includes built-in text analysis functionalities, such as quick indexing of large collections of documents, embedded natural language processing (sentence splitting, tokenization and shallow parsing) and querying the text collection using weighted fuzzy and linguistic matching. Our evaluation for the passage retrieval task has shown the particularities of each language and has pointed out which resources are necessary for each of them in order to boost multilingual question answering results for the biomedical domain.

7. Acknowledgements

We would like to thank our colleagues in the Hasso-Plattner Institut (HPI), specially Dr. Matthieu-P. Schapra-

[13] http://dbpedia.org/

now, Cindy Fähnrich and Thomas Uhde. MN would like to acknowledge funding from the HPI Research School.

8. References

Sofia J. Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1 – 24.

MichaelA Bauer and Daniel Berleant. 2012. Usability survey of biomedical question answering systems. *Human Genomics*, 6(1):1–4.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 609–617, Stroudsburg, PA, USA. Association for Computational Linguistics.

William Hersh and Ellen Voorhees. 2009. Trec genomics special issue overview. *Inf. Retr.*, 12(1):1–15, February.

Antonio Jimeno Yepes, Elise Prieur-Gaston, and Aurelie Neveol. 2013. Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. *BMC Bioinformatics*, 14(1):146.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante, Martin Krallinger, Alfonso Valencia, and Walter Daelemans. 2012. Machine reading of biomedical texts about alzheimer's disease. In *CLEF (Online Working Notes/Labs/Workshop)*.

Mara-Dolores Olvera-Lobo and Juncal Gutirrez-Artacho. 2011. Multilingual question-answering system in biomedical domain on the web: An evaluation. In Pamela Forner, Julio Gonzalo, Jaana Keklinen, Mounia Lalmas, and Maarten de Rijke, editors, *CLEF*, volume 6941 of *Lecture Notes in Computer Science*, pages 83–88. Springer.

Ioannis Partalas, Eric Gaussier, and Axel-Cyrille Ngonga Ngomo. 2013. Results of the first bioasq workshop. In *1st Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013)*, nov.

Dietrich Rebholz-Schuhmann, Simon Clematide, Fabio Rinaldi, Erik Van Mulligen, Ian Lewin, David Milward, Antonio Jimeno Yepes, Udo Hahn, Jan Kors, Chinh Bui, Johannes Hellrich, and Michael Poprat. 2013. Multilingual semantic resources and parallel corpora in the biomedical domain: the clef-er challenge. In *CLEF Lab*, September.

Matthieu-P. Schapranow, Hasso Plattner, and Christoph Meinel. 2013. Applied in-memory technology for high-throughput genome data processing and real-time analysis. In *System on Chip (SoC) Devices in Telemedicine from LABoC to High Resolution Images, pp. 35-42, ISBN: 978-84-615-6080-6*.

Resolving Abbreviations in Clinical Texts Without Pre-existing Structured Resources

Borbála Siklósi, Attila Novák, Gábor Prószéky
Pázmány Péter Catholic University, Faculty of Information Technology and Bionics,
MTA-PPKE Hungarian Language Technology Research Group
50/a Práter street, 1083 Budapest, Hungary
{surname.firstname}@itk.ppke.hu

Abstract

One of the most important topics in clinical text processing is the identification of relevant concepts. This includes the detection and resolution of abbreviations, acronyms, or other shortened forms in the documents. Even though the task of resolving abbreviations can be treated as a word sense disambiguation problem, such methods require structured lexical knowledge bases. However, for less-resourced languages such resources are not available. In this paper, a method is proposed for the disambiguation and resolution of abbreviations found in Hungarian clinical records. In order to achieve reasonable performance, a lexicon must be created for each domain. It is shown how the set of entries to be included in such a lexicon can be induced from the corpus, thus the manual effort of creating a lexicon can be reduced significantly. The results for resolving abbreviations in Hungarian clinical documents are also shown, which are achieved by using the corpus instead of non-existing structured resources.

Keywords: clinical NLP, abbreviation resolution, less-resourced languages

1. Introduction

Processing medical texts is an emerging topic in natural language processing. There are existing solutions mainly for English to extract knowledge from medical documents, which will be available for researchers and medical experts. However, locally relevant characteristics of applied medical protocols or information relevant to locally prevailing epidemic data can be extracted only from documents written in the language of the local community. In the case of less-resourced languages, such as Hungarian, the lack of structured resources, like UMLS (Lindberg et al., 1993), Snomed (International Health Terminology Standards Development Organisation, 2010), etc. makes it very hard to produce results comparable to those achieved by solutions for major languages. One way to overcome this problem could be the translation of these resources, however, doing it manually would require a huge amount of work, and automated methods that could support the translation effort are also of low quality for these languages.

Beside the availability of structured resources, categories of medical text processing can also be differentiated according to the type of text being processed (Meystre et al., 2008). Most research focuses on biomedical texts that appear in books, articles, literature, etc. However, there are a growing number of studies on processing clinical texts written by doctors in the clinical settings. This paper fits this latter line of research.

In Hungarian hospitals, clinical records are created as unstructured texts without using any proofing tools, resulting in texts full of spelling errors and nonstandard use of word forms in a language that is usually a mixture of Hungarian and Latin (Siklósi et al., 2012; Siklósi et al., 2013). These texts are also characterized by a high ratio of abbreviated forms. The use of some of these abbreviations follows some standard rules, but most of them are used in an arbitrary manner. Moreover, in most cases, full statements are written in a special notational language (Barrows et al., 2000) that is often used in clinical settings, consisting only, or mostly of abbreviated forms. Even for non-expert humans, it is a hard task to find phrase boundaries in a long sequence of shortened forms. Processing such documents is not an easy task, and resolving abbreviations is a prerequisite of further linguistic processing.

The task of abbreviation resolution is often treated as word sense disambiguation (WSD) (Navigli, 2012). The best-performing approaches of WSD use supervised machine learning techniques. In the case of less-resourced languages, however, neither manually annotated data, nor an inventory of possible senses of abbreviations are available, which are prerequisites of supervised algorithms (Nasiruddin, 2013). On the other hand, unsupervised WSD methods are composed of two phases: word sense induction (WSI) must precede the disambiguation process. Possible senses for words or abbreviations can be induced from a corpus based on contextual features. However, such methods require large corpora to work properly, especially if the ratio of ambiguous terms and abbreviations is as high as in the case of clinical texts. Due to confidentiality issues and quality problems, this approach is not promising either.

In this study, we introduce the behaviour of abbreviations in clinical documents of low-resourced languages, demonstrated with our Hungarian corpus of medical records. Then, a corpus-based approach is described for the resolution of abbreviations with using the very few lexical resources available in Hungarian. As this method did not provide acceptable results, the construction of a domain-specific lexicon was unavoidable. Instead of trying to create huge resources covering the whole field of medical expressions, it is shown that small domain-specific lexicons are satisfactory and the abbreviations to be included can be derived from the corpus itself. Finally, an analysis of the combination of these methods is presented.

2. Related work

There are several studies, most of them applied to English texts, that address the specific task of medical language processing. One of the most challenging preprocessing steps is the detection and resolution of abbreviations found in the free-text parts of these documents. As opposed to biomedical literature, where the first mention of an abbreviated form is usually preceded by its expanded form or definition, in clinical records this is not the case. That is why simple abbreviation-definition patterns are not applicable to clinical notes as described by Xu et al. in (Xu et al., 2007). The same study compares some machine learning approaches, all achieving considerable results, but even using already existing external resources, the authors admit the need of a manually created inventory. A recent study (Wu et al., 2012) compared the performance of some biomedical text processing systems trained on biomedical literature on the task of resolving abbreviations in clinical texts. All the systems (MetaMap, MedLEE and cTAKES) achieved suboptimal results calling for more advanced abbreviation recognition modules.

Most approaches to resolving clinical abbreviations carried out in English rely on some very common medical lexical resources. Even though Xu in (Xu et al., 2007) showed that the sense inventories generated from the UMLS covered only about 35% of the abbreviations they had extracted from their corpus, these already contain definitions and possible interpretation candidates. Thus the problem can be reduced to abbreviation disambiguation, as it is carried out in (Wu et al., 2012; Xu et al., 2009; Pakhomov et al., 2005). These methods focus primarily on supervised machine learning approaches, where a part of the training corpus is labeled manually. Pakhomov in (Pakhomov, 2002) described a semi-supervised method to build training data for Maximum Entropy modeling of abbreviations automatically. In most of these studies, both training and evaluation of the systems are performed on a few manually chosen abbreviations and their disambiguation.

3. Clinical abbreviations

The use of a kind of notational text is very common in clinical documents. This dense form of documentation contains a high ratio of standard or arbitrary abbreviations and symbols, some of which may be specific to a special domain or even to a doctor or administrator. These shortened forms might refer to clinically relevant concepts or to some common phrases that are very frequent in the specific domain. For the clinicians, the meaning of most of these common phrases is as trivial as the standard shortened forms of clinical concepts due to their expertise and familiarity with the context. Some examples for abbreviations falling into these categories are shown in Table 1.

3.1. Series of abbreviations

Even though standalone abbreviated tokens are highly ambiguous, they more frequently occur as members of multiword abbreviated phrases, in which they are usually easier to interpret unambiguously. For example *o.* could stand for any word either in Hungarian or in Latin, starting with the letter *o*, even if limited to the medical domain. However,

Domain	Abbr.	Resolution	in Hungarian	in English
standard	o. d. med. gr.	oculus dexter mediocris gradus	jobb szem közepes fokú	right eye medium grade
domain-specific	o. o.	oculus os	szem csont	eye bone
domain-specific common	sü fén n	saját szemüveg fényérzés nélkül normál	saját szemüveg fényérzés nélkül normál	own glasses no sense of light normal
common words	köv lsd	következő lásd	következő lásd	next see

Table 1: Some examples for the use of simple abbreviations. Some of them are commonly known standard forms, usually of Latin origin, some others, though related to the clinical domain, might have several meanings depending on the specific sub-domain. The rest are abbreviated common words, usually of Hungarian origin, and might also refer to both clinical phrases or common words.

in our corpus of anonymized ophthalmology reports, *o.* is barely used by itself, but together with a laterality indicator, i.e. in forms such as *o. s., o. d.*, or *o. u.* meaning *oculus sinister* 'left eye', *oculus dexter* 'right eye', or *oculi utriusque* 'both eyes', respectively. In such contexts, the meaning of the abbreviated *o.* is unambiguous. It should be noted, that these are not the only representations for these abbreviated phrases, for example *oculus sinister* is also abbreviated as *o. sin., os, OS*, etc. Table 2 shows the ratio of unique abbreviation sequences of different lengths detected automatically in the corpus with the method described in Section 5.1. The number of different single-token abbreviations is roughly equal to the number of all multi-token abbreviations.

Length:	1	2	3	4	5	>5
Number:	49.53%	26.34%	15.00%	5.95%	2.16%	0.98%

Table 2: The ratio of unique abbreviation series of different lengths detected automatically in the corpus.

Thus, when performing the resolution of abbreviations, we considered series of such shortened forms instead of single tokens. A series is defined as a continuous sequence of shortened forms without any unabbreviated word breaking the sequence. These series are not necessarily coherent phrases. The individual elements of such sequences of abbreviations are by themselves highly ambiguous, and even if there were an inventory of Hungarian medical abbreviations, which does not exist, their resolution could not be solved. Moreover, the mixed use of Hungarian and Latin phrases results in abbreviated forms of words in both languages, thus the detection of the language of the abbreviation is another problem. For example, in the "sentence"

Dg : Tu. pp. inf et orbitae l. dex. ,
Cataracta incip. o. utr. , Hypertonia,

the abbreviation spans are the following:

Dg,
Tu. pp. inf,
l. dex.,
incip. o. utr..

3.2. The lexical context of abbreviation sequences

In the above example, the last section is misleading, since the token *incip.* is part of the phrase *Cataracta incip.*, i.e. it is related to its preceding neighbour, which is not included in this list as part of an abbreviation. This mixed use of a phrase is very common in the documents, with a diverse variation in using certain words in their full form or in some shortened form instead. In order to save such phrases and to keep the information relevant for the resolution of multiword abbreviations, the context of a certain length is attached to the detected series. In our experiments, the length of the context taken from both the left and right sides of the abbreviations ranged from 0 to 3 tokens. Since the average length of sentences in the corpus is 9.7 (Orosz et al., 2013), considering a larger context could span across sentences, but that would make no sense.

Beside completing such mixed phrases, the context also plays a role in the process of disambiguation. The meaning (i.e. the resolution) of abbreviations of the same surface form might vary in different contexts. Our experimental results showed that this does not require a larger window of sampling either.

4. Resources

The corpus In our research, we used a corpus of anonymized clinical documents, all falling into the domain of ophthalmology. A portion of this corpus was set aside for testing purposes. Table 3 contains detailed information about the size of these subcorpora.

	documents	sentences	tokens	abbreviated tokens
whole corpus	2008	60660	552594	113091
test corpus	22	693	5599	765

Table 3: The size of our corpus of clinical records

External lexicon Even though there are no structured lexical resources for Hungarian, the official coding system for diseases, anatomical structures and medical procedures is available (similar to the ICD systems in English). Thus, a simple dictionary was built from the ophthalmology sections of these descriptions. The final list of phrases contained 3329 entries. However, these phrases are written in the language of the official terminology, which is different in several respects from that used in the clinical texts.

Handmade lexicon Since the official descriptions turned out not to be of much use, a domain-specific lexicon seemed to be necessary. The first step of designing such a resource is to decide what phrases to include. We assumed that the most frequent abbreviations occurring in the corpus without their expanded form ever being written out have one unambiguous resolution within a narrow domain. For example, in the domain of ophthalmology, the abbreviation *o. d.* always stands for the phrase *oculus dexter*, meaning 'right eye'. Even though it appears in various shortened forms, it is never spelled out in its full form. Thus, a frequency list of abbreviations and abbreviation series was created from the whole corpus. Then a threshold value was defined experimentally and the abbreviations with a relative frequency above this threshold were included into our set of domain-specific abbreviations. Finally, the resolution of these abbreviations were defined with the help of a medical expert.

This approach can be applied to other domains as well. Thus, our method of abbreviation resolution is applicable to new domains with a relatively small amount of manual effort. In our case of ophthalmology records, rather good coverage can be achieved even with a small lexicon of 44 entries. Adding more items to the list further improves the quality of the resolution, however, the improvement achieved by adding new items diminishes quickly.

5. Methods

The primary objective of the research described here is finding spans in a sequence of abbreviations that can be unambiguously resolved together. Since sometimes whole statements or even sentences are written using this kind of heavily abbreviated notation, it is important to find an optimal partitioning of the tokens into meaningful spans. In the previous example, the fragment *incip. o. utr.* should be divided into the spans of *incip.* and *o. utr.*, even if the abbreviation *incip.* is not relevant by itself, but still its meaning is not related to the rest of the abbreviation sequence. However *o. utr.* can be resolved with high confidence.

5.1. Detection of abbreviations

The first problem to solve when trying to handle abbreviations within running text is detecting them. Since these texts usually do not follow standard orthographic and punctuation rules, especially in the case of highly abbreviated notational text, the detection of abbreviations cannot be based on patterns formulated according to standard rules of forming abbreviations. The ending periods are usually missing, abbreviations are written with varying case (capitalization) and in varying length. For example the following forms represent the same expression, *vörös visszfény* 'red reflection': *vvf, vvfény, vörösvfény*. We applied some heuristic rules to derive relevant features as indicators of a token being an abbreviation. These features were based on the following characteristics: the presence or absence of a word-final period, the length of the token, the ratio of vowels and consonants within the token, the ratio of upper- and lowercase letters, and the judgment of a Hungarian morphological analyzer, the lexicon of which was expanded with medical terminology (Novák, 2003; Prószéky and Kis, 1999).

5.2. Resolving abbreviations

Once the abbreviation series are extracted from a document, a maximum coverage resolution suggestion process is carried out. The steps of the algorithm are the following:

1. For each possible partitioning of the tokens into non-overlapping spans, a regular expression pattern is generated. The patterns are created by general abbreviation rules, such that each letter in the abbreviated form represents the starting letter of each word in the expanded phrase (assuming that it is an acronym). Or,

in the case of multiword abbreviations, each member represents the beginning of each word (not just the first letter) in the interpretation. Some pattern generation rules are presented in Table 4.

2. The regular expressions for each span are matched against the corpus resulting in full or partial resolutions.

3. The regular expressions are refreshed based on the results retrieved from the corpus.

4. These expressions are then matched against the lexicons.

5. The results of the spans are concatenated to cover the whole series and each such merged resolution candidates are given a score appropriate for ranking.

6. The highest ranked resolution is considered as the final one.

During this process, the optimal division of the abbreviation series and its resolution are carried out in one step. This is ensured by the scoring method. The different partitionings are ranked according to three features, optimizing for the longest coverage and best resolution of the abbreviation sequence. The features used for ranking are 1) the number of all tokens in the sequence covered by a resolved form, 2) the size of the longest span covered, 3) the size of the shortest span covered. If the sequence *Exstirp. tu. et reconstr. pp. inf. l. d.*, is partitioned as | *Exstirp. tu.* | *et* | *reconstr. pp. inf.* | *l. d.* |, with having a resolution candidate for each partition except for the word *et*, then the value of the features are 7, 3 and 2, respectively. Finally, these values given for each feature are transformed into a percentage score by a weighted combination of them.

5.3. Experiments

The algorithm described above uses three resources where the resolution candidates are searched for: 1) a lexicon containing the ophthalmology section of the official ICD coding system, 2) our handmade lexicon, and 3) the corpus itself. In our experiments, we investigated how the performance of the algorithm is influenced by the availability of these resources to the program. The goal of these experiments were threefold. First, since there is a lack of structured external resources, we wanted to investigate to what extent we could rely on a raw, domain-specific corpus to resolve abbreviations. In order to do this, the size of the corpus in which the regular expressions were matched was changed incrementally. Second, we wanted to check the hypothesis that a small, manually built lexicon (containing the most frequent abbreviations) can be built and utilized and in an effective manner. Moreover, we wanted to identify a threshold for the collected entry candidates to such a lexicon for an arbitrary domain. Third, the best performing combination was evaluated.

6. Results and discussion

A test set of 22 documents was used for evaluation purposes both for the task of abbreviation detection and resolution.

abbr	regexp	matching expansion
o. s.	`o[^ ]* s[^ ]*`	oculus sinister
os	`os[^ ]*`	osteoporosis
os	`o[^ ]* s[^ ]*`	oculus sinister

Table 4: Some of the simplest patterns generated from two short abbreviated phrases. The complexity and variability of these patterns is proportional to the length of the original abbreviation sequence.

	result
Precision:	95.99%
Recall:	97.12%
F-measure:	96.55%

Table 5: Evaluation results for abbreviation detection

The abbreviations in this set were labeled manually and resolved by a medical expert. Finally, the number of tokens labeled as abbreviations was 765. The actual meaning for 56 of them could not be specified. These included initials of doctors; author-specific shortened forms; tokenization errors, etc. Thus, the remaining 709 abbreviations were considered when evaluating our methods for abbreviation resolution.

Performance in both tasks was measured in terms of precision, recall and F-measure. For abbreviation detection, precision was calculated as the number of true positives divided by the sum of true positives and false positives and recall was defined as the number of true positives divided by the sum of true positives and false negatives. On the other hand, for the resolution task, precision was defined as the number of correctly resolved tokens divided by the number of all resolved tokens, while recall is the number of correctly resolved tokens divided by the number of all abbreviations (Cohn, 2003). Thus, if having one abbreviation resolved correctly without touching anything else, a precision of 100% could be achieved, which does not reflect the real performance. That is why F_2-measure was defined as the harmonic mean between precision and recall biased towards recall.

Table 5 shows the final results for the detection of abbreviations. Most of the errors in the detection arose from misspelled forms, tokenization errors or Latin abbreviations.

In the case of abbreviation resolution, the effect of varying four parameters were investigated. The first parameter was the size of the context taken into consideration when resolving abbreviation sequences. Second, we investigated the effect of changing the size of the corpus used for pattern matching. Third, the effect of changing the size of our handmade lexicon and finding the optimal threshold value to decide what to include in it. And fourth, the performance of the system using the best combination of these parameters was evaluated.

Figure 1 shows the results of three experimental setups ((a),(b) and (c)). In each of them, the size of the manually created lexicon was kept at a fixed size (0, 44, and 136 entries). The size of the corpus was increased in units of

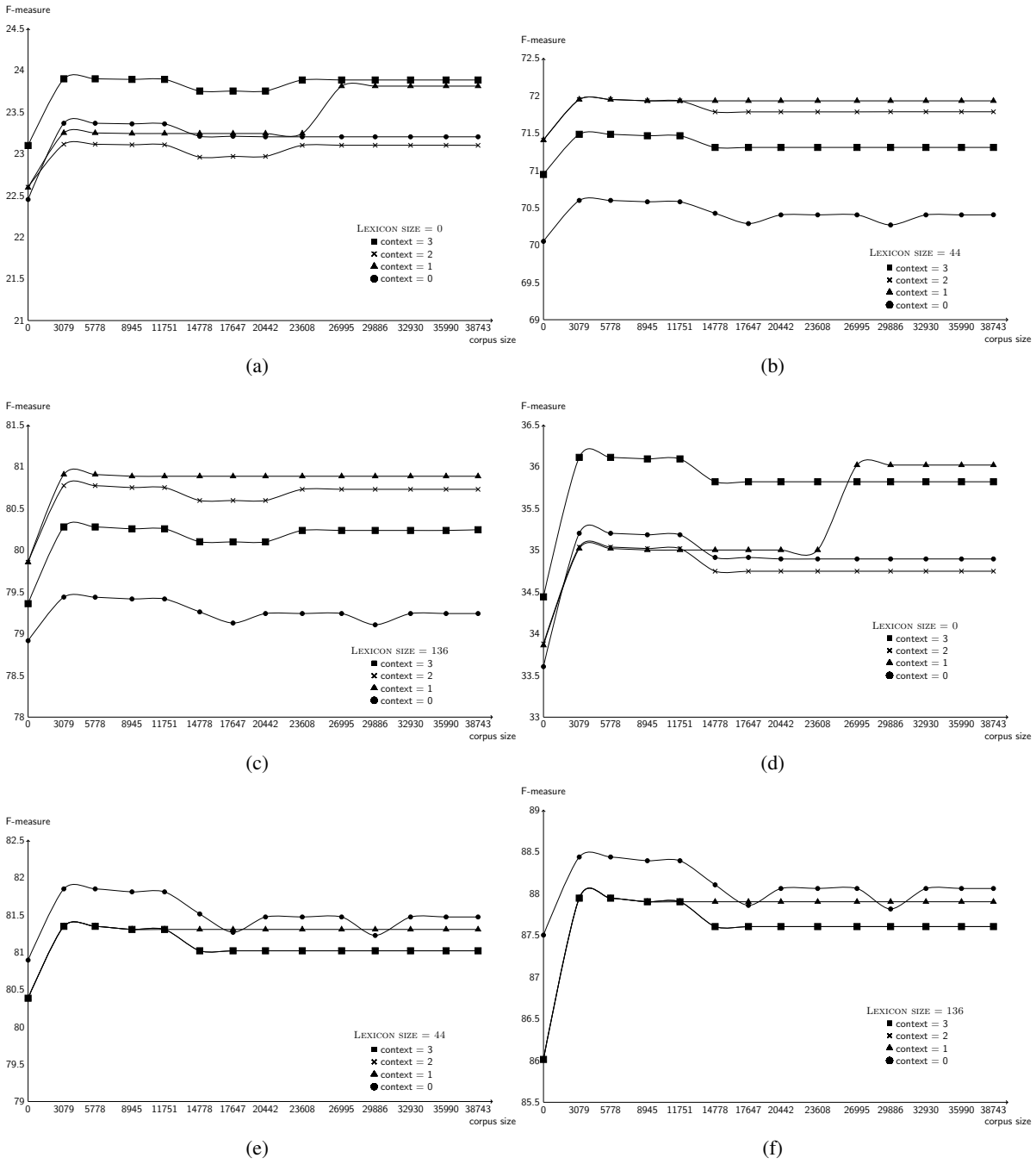


Figure 1: The performance results as a function of the corpus size for different context sizes and using a fixed portion of our handmade lexicon (0, 44 and 136 entries respectively). Graphs (a), (b) and (c) represent the results for all abbreviation series, while graphs (d), (e) and (f) represent the results for multi-token abbreviation sequences only.

around 3000 sentences in each step and the performance for context sizes 0 to 3 tokens were measured. Replacing each abbreviation with its definition from the lexicon (if it was included in the lexicon) was considered as the baseline. In the first case, without the lexicon, this baseline was an F-measure of 0%, in the second case 60.52%, and in the third case, it was 75.71%. As it can be seen from the graphs, these values are quite below the performance of our final combined system. In each case, considering the tokens without any context always performs worst, however, taking a context larger than one token before and after the abbreviation has a positive effect only if the manually created lexicon is not used. In this case, the system with a context size of three tokens performed best.

Increasing the size of the corpus had a similar effect. When the domain-specific lexicon is available, then the only significant change in performance occurred when adding the first portion of the corpus. Further increasing its size did not influence the performance of these setups. This is due to the relatively small size of the corpus and the noisy nature of the texts.

In (Siklósi and Novák, 2013), it was reported that by in-

creasing the size of the corpus, the difference between the performance achieved by using lexicons of different sizes can be made up. However, in that study, abbreviations of multiple tokens were considered (i.e. no single abbreviated tokens) and the test set contained unique abbreviations. Thus we also performed the evaluation tests on abbreviation sequences of length greater than 1 (see graphs (d), (e) and (f) of Figure 1). Comparing the behaviour of the algorithm for all and for multiple-token abbreviations, there are two main differences. First, the performance values are higher for longer series of abbreviations. Second, taking a context of any size performs worse than having only the abbreviation by itself in the case of such longer series. Moreover, we found that, when using the lexicon, adding too much of the corpus will generate noise for the resolution process instead of enhancing the quality.

In order to find an optimal relative occurrence frequency threshold value for abbreviations that should be included in the handmade lexicon, the largest corpus size was used and the abbreviations were added to the lexicon incrementally. Figure 2 shows the change in the threshold and the performance as a function of the number of entries in the lexicon. The results are in accordance with our assumption that including the most frequent abbreviations in the lexicon has a more significant role than creating a large, more detailed lexicon. Adding only the first 10 most frequent abbreviations to the lexicon results in a 30% increase in the performance. Even though the performance grows further if adding more entries, an optimal threshold can be set by cutting the long tail of the graphs. Thus, in our case, this cut was done where abbreviations with relative frequency higher than 0.0025 were added to the lexicon. It resulted in a lexicon size of 44 entries.

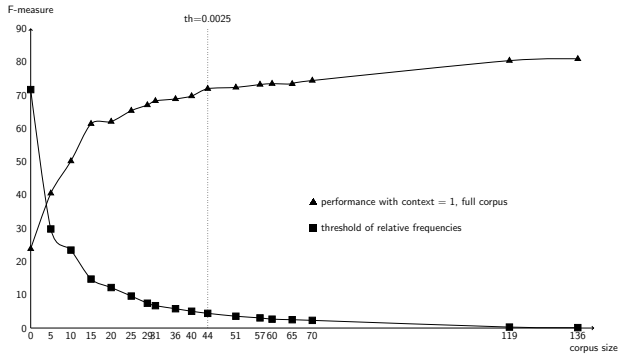


Figure 2: The change in the threshold and the performance as a function of the number of entries in the lexicon. Decreasing the threshold (measured in relative corpus frequency) below the value of 0.0025 does not produce a significant increase in the performance relative to the manual effort needed to define the meaning of these abbreviations. The F-measure values here correspond to a context size of one token and the whole corpus is used for pattern matching.

Even though the above investigations are important in order to be able to generalize the system to other domains or languages, and filling the gap caused by the lack of structured resources, our goal was also to achieve the best results in resolving abbreviations in Hungarian clinical records in the domain of ophthalmology. The best results compared to those achieved by using the above described threshold are shown in Table 6. In the case of the final setup, pattern matching was applied to the whole corpus with taking a one-token context around each abbreviation, and using an enlarged version of our lexicon to 136 entries. (We have no data about further increasing this size.) Thus, an F-measure of 80.88% was achieved for all abbreviations and 88.05% for abbreviation series consisting of multiple tokens.

length	lexicon size	precision	recall	f2
all	136	93.23%	78.29%	80.88%
	44	88.37%	68.73%	71.92%
>1	136	96.22%	86.23%	88.05%
	44	90.63%	79.46%	81.46%

Table 6: The best performance achieved for the ophthalmology corpus for all abbreviations and for abbreviation series of length greater than 1.

7. Conclusion

Automatic detection and resolution of abbreviations in clinical documents are usually solved by using external resources. In this study an approach was presented to solve the same problems if such resources are not available, which is the case for less-resourced languages, such as Hungarian. It has been shown that the presence of a domain-specific lexicon is crucial, however it does not need to be a large, detailed knowledgebase. A small lexicon can be created by defining the resolution for the most frequent abbreviations found in a corpus of a narrow domain. The rest of the abbreviations can be resolved based on the corpus itself. On the other hand, the role of the context of an abbreviated token was investigated from different aspects. It has been shown that ambiguous abbreviations are much easier to be interpreted as members of abbreviation series, moreover, adding a one token long context to these series has also beneficial effect on the performance of the disambiguation process.

8. Acknowledgement

The authors of this paper would like to thank Ákos Kusnyerik for providing his expertise knowledge in ophthalmology.

This research was partially supported by the project grants TÁMOP–4.2.1./B–11/2-KMR-2011-0002 and TÁMOP–4.2.2./B–10/1-2010-0014.

9. References

Barrows, J. R., Busuioc, M., and Friedman, C. (2000). Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Proceedings of the AMIA Annual Symposium*, pages 51–55.

Cohn, T. (2003). Performance metrics for word sense disambiguation. In *Proceedings of the Australasian Language Technology Workshop (ALTW)*, pages 49–56, Melbourne, Australia, December.

International Health Terminology Standards Development Organisation. (2010). Snomedct: Systematized nomenclature of medicine-clinical terms.

Lindberg, D., Humphreys, B., and McCray, A. (1993). The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291.

Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44.

Nasiruddin, M. (2013). A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages. *CoRR*, abs/1310.1425.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pages 115–129.

Novák, A. (2003). What is good Humor like? In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged. SZTE.

Orosz, G., Novák, A., and Prószéky, G., (2013). *Hybrid text segmentation for Hungarian clinical records*, volume 8265 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, Heidelberg.

Pakhomov, S., Pedersen, T., and Chute, C. (2005). Abbreviation and acronym disambiguation in clinical discourse. In Friedman, C., Ash, J., and Tarczy-Hornoch, P., editors, *Proceedings of the AMIA Annual Symposium*, pages 589–593, Washington DC, USA. Bethesda, MD: AMIA Press.

Pakhomov, S. (2002). Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In Isabelle, P., editor, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 160–167, Philadelphia, USA. Rochester, NY: ACL Press.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 261–268, Stroudsburg, PA, USA. Association for Computational Linguistics.

Siklósi, B. and Novák, A., (2013). *Detection and Expansion of Abbreviations in Hungarian Clinical Notes*, volume 8265 of *Lecture Notes in Artificial Intelligence*, pages 318–328. Springer-Verlag, Heidelberg.

Siklósi, B., Orosz, G., Novák, A., and Prószéky, G. (2012). Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., De Schryver, G.-M., Forcada, M., M. Tyers, F., and Waiganjo Wagacha, P., editors, *8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*, pages 29.–34.

Siklósi, B., Novák, A., and Prószéky, G. (2013). Context-aware correction of spelling errors in Hungarian medical documents. In Dediu, A.-H., Martin-Vide, C., Mitkov, R., and Truthe, B., editors, *Statistical Language and Speech Processing*, volume LNAI 7978. Springer Verlag.

Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A., and Xu, H. (2012). A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. *Proceedings of the AMIA Annual Symposium*, 2012:997–1003.

Xu, H., Stetson, P., and Friedman, C. (2007). A study of abbreviations in clinical notes. In Teich, J., Suermondt, J., and Hripcsak, G., editors, *Proceedings of the AMIA Annual Symposium*, pages 821–825, Washington DC, USA. Bethesda, MD: AMIA Press.

Xu, H., Stetson, P., and Friedman, C. (2009). Methods for building sense inventories of abbreviations in clinical notes. *Journal of American Medical Informatics Association*, 16(1):103–108.

Worlds Apart – Ontological Knowledge in Question Answering for Patients

Lina Henriksen, Anders Johannsen, Bart Jongejan, Bente Maegaard, Jürgen Wedekind
University of Copenhagen, Center for Language Technology
Njalsgade 140, 2300 Copenhagen S, Denmark
{linah, ajohannsen, bartj, bmaegaard, jwedekind}@hum.ku.dk

Abstract

We present ESICT, a hybrid question-answering system building on formalized knowledge from a medical ontology (SNOMED CT) as well as text-to-text generation in the form of document summarization and question generation. The independent subsystems are queried in parallel and compete for delivering the best answer. The use of an ontology gives the patient access to information typically not found in other sources, but also exposes a gap between everyday language and the specialized terms and conceptualizations of health professionals. In this paper we describe the ESICT system and discuss for each strategy how well it deals with this gap.

Keywords: question answering, nlidb, eHealth

1. Introduction

There is an emerging focus on the active citizen taking responsibility for his own health and illness through actively seeking and using information. But two people suffering from the same disease and seeking the same information might not ask questions in the same way and might not find the same answers equally useful. Citizens have different backgrounds in terms of education, social circumstances, new diagnoses vs. long-term chronic disease histories, etc.—issues that put very heavy demands on eHealth systems.

In Denmark a number of eHealth information systems are available on the Internet such as *netdoktor.dk* and *sundhed.dk*. However, they are to a large extent based on Frequently Asked Questions, they contain text chunks in a flat structure, and their purpose is to provide the citizen with an overview of predefined topics. Question-answering (QA) technology is another approach to provide citizens with quick and easy access to health and disease related information.

Existing QA systems have different strengths and limitations (cf., e.g., Athenikos and Han, 2010, for a survey). Those based on text-to-text generation are limited by the relevance and accuracy of the underlying text collections. Among the strengths of these systems are that the underlying text collection can be very large and the retrieved answer's wording and style will typically reflect the question and therefore be intelligible to the user. QA systems based on terminologies, ontologies, and other kinds of formalized knowledge often command clear, unambiguous, and terminologically correct information that may, however, be very different from the words used and known by the user. Besides, these systems require at least a shallow understanding of the user's question in order to provide a meaningful answer; recognition of a few keywords is not sufficient. The question *Can diabetes result in blindness?* is different from *Will diabetes result in blindness?* and *Can blindness result in diabetes?* and they require different answers. Further, such systems provide facts and not linguistically well-formed answers, and often the user's question is the best choice as a starting point for the wording of an answer.

We present ESICT (Experience oriented Sharing of health knowledge via Information and Communication Technology) (Andersen et al., 2012), a hybrid QA system employing highly structured as well as less structured information and offering information about diabetes mellitus in Danish. The coverage of ESICT is based on a corpus of 321 diabetes questions collected from three different sources: (i) an online diabetes discussion forum; (ii) an outpatient clinic at a Danish hospital (Wizard of Oz sessions); and (iii) a workshop with health informatics students. These real-life questions collected from patients, relatives, and other citizens reflect many complexities such as modality as in *Can diabetes be hereditary?* ambiguity as in *Is diabetes curable?* and personal issues as in *Can I get blood clots from diabetes?* or *Can I eat chocolate?* Most questions are within the topics: molecular and biomedical facts, epidemiology, interventions (e.g., behavioral intervention in terms of diet, exercise and other life style issues), and diagnostics.

ESICT applies three different approaches to question processing and answer generation. One approach relies on SNOMED CT, a multilingual clinical healthcare terminology covering terms of anatomy, findings, procedures, etc. in sub-type (is-a) hierarchies supported by defining relationships.[1] SNOMED CT is considered the world's most comprehensive nomenclature of clinical medicine. Examples of computer applications using SNOMED CT include electronic patient journal systems, clinical decision support systems, laboratory reporting systems, and many more.[2] Because SNOMED CT, with its hierarchical design and primarily definitional knowledge, only covers some types of user questions, we also investigated two alternative QA strategies that could backup or, in cases of failure, even replace SNOMED CT-based querying. These are text-to-text generation approaches that both draw on authoritative medical texts within the diabetes domain. One approach uses query-focused multi-document summarization whereas the other generates potential users' questions on the basis of the particular document collection. These approaches all work in parallel in a running prototype. The following sections include information about the status, challenges, and

[1] http://www.ihtsdo.org/snomed-ct
[2] http://www.ncbi.lm.nih.gov/pmc/articles/PMC3704061

perspectives of each approach.
In this paper we will also discuss the pros and cons of each approach with respect to coherence between question and answer in terms of style, word selection, accuracy, etc. Another aspect which will be discussed is an inherent difficulty of medical QA systems: Not only do laymen phrase questions in different ways, as mentioned in the introduction, but laymen and health professionals have very different conceptual models of the world (cf., e.g., Zhang, 2010). In this paper we will try to discuss this problem for each of the approaches and evaluate their capacity for bridging the gap between the layman's and the doctor's conceptual worlds.
The rest of the paper is organized as follows. Sections 2–4 describe the three strategies of the ESICT QA system: ontology-based QA, multi-document summarization, and question generation. Section 5 presents the evaluation of the ESICT system. The last section highlights our key observations.

2. Strategy A: Ontology-based QA

The main focus of this approach was to build a natural language interface to SNOMED CT. We transform natural language questions into queries on the SNOMED CT ontology and produce natural language answers based on the results of these queries. Both the natural language questions and their answers are systematically related to SNOMED CT interpretable expressions, in the following called SNOMED expressions. SNOMED expressions are composed of atomic relational statements (triplets of the form $concept_1$ – *relation* – $concept_2$) and the description-logical operators supported by SNOMED CT. (In the actual implementation SNOMED expressions, together with the dependency analyses of the questions, are compiled into ontological scripts enabling SNOMED CT to infer the requested information.)

2.1. Workflow

The user's question in natural language is mapped onto a SNOMED expression by rewriting the dependency analysis of the question. The dependency analysis is created through a chain of natural language processing tools.[3] Rewriting is accomplished by a set of transformation rules[4] and results in a semantic analysis of the question. For example, the question *Får man katarakt af diabetes?* (Do you get cataract from diabetes?) is analyzed and transformed into a SNOMED expression as illustrated in Figure 1. The figure shows the relation between the decomposed syntactic analysis and the corresponding SNOMED CT terms. The area delineated with a red line is mapped onto the SNOMED CT relation DUE TO, while the blue delineated areas are mapped onto SNOMED CT concepts. In many questions the relation is expressed by a transitive main verb and the

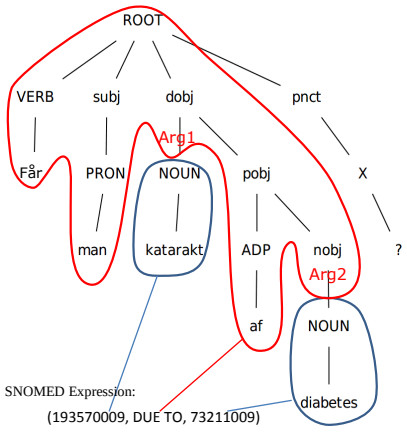


Figure 1: Dependency parse tree and semantic analysis of the question *Får man katarakt af diabetes?*

concepts are expressed by the subject and the object. However, there are also many other questions, such as predicative constructions (e.g., *Is diabetes type 2 dangerous?*) or questions with ditransitive verbs (such as in Figure 1 with non-pronominal subjects) that require a more sophisticated analysis.
Each SNOMED CT relation accepts only arguments of certain types. In case of the DUE TO relation these types are CLINICAL FINDING and EVENT for both the first and the second argument. These restrictions are context constraints on the application of the mapping rules and are used to disambiguate the semantic interpretation of the dependency structures.
The dependency parse tree is not only used as input for the semantic analysis but also as the raw material for the preparation of the answer. By replacing WH-phrases and pronouns, inserting function words, and moving constituents, the interrogative form of the user's utterance is transformed into a linguistically well-formed declarative answer.

2.2. Strengths and limitations

SNOMED CT has been developed using a variant of the relatively inexpressive description logic $\mathcal{EL}$. Thus it provides only very limited semantic expressivity and reasoning support. Modal, causal, and temporal reasoning, quantifier scoping, negations, comparatives, superlatives, etc. cannot be properly accounted for within an ontology like SNOMED CT. Therefore, for example, procedural how questions (e.g., *How do I prevent eye damage in diabetes?*), explanation questions (e.g., *Why does diabetes affect the liver?*), and advice-seeking questions (e.g., *How do I cope with diabetes?*) are generally out of scope of strategy A.
There are also many questions on topics not dealt with by the ontology. For instance, a diabetes patient might ask whether a particular food item is compatible with his illness and if not under what circumstances allowances may be made (e.g., *Can I eat pork rinds at the Christmas dinner?*). This is an out-of-topic question because diet recommendations are not included in SNOMED CT.
Our generation component is currently dependent on the input analysis. Thus we cannot generate answers that are structurally unrelated to the question (as, e.g., more elab-

[3] Consisting of a simple tokenizer, the OpenNLP tagger (http://opennlp.apache.org/), Bohnet's parser (http://code.google.com/p/mate-tools/), and CST's lemmatizer (Jongejan and Dalianis, 2009).

[4] Expressed as a script in a tree pattern matching language, Bracmat (https://github.com/BartJongejan/Bracmat).

orate answers to the question *Which treatments are available for diabetes?*). However, there are dependency-based stand-alone generators (e.g., Guo et al., 2011) that can, in combination with our mapping rules, be adapted to generate natural language expressions from SNOMED CT expressions without reference to the input analysis.

The presence of vagueness and ambiguity in ordinary language concepts presents an enormous challenge for disambiguating and interpreting layman's health-related questions in the SNOMED CT ontology. This is because there is usually a multi-to-one correspondence between natural language concepts and the medical concepts of SNOMED CT, and vice versa. This problem is exacerbated by SNOMED CT's relational sparseness (there is only a rather small number of relation types). Thus, there are usually numerous predicates that map to the same relation, and there are also many predicates that map onto different relations, depending on their arguments. This rather loose correspondence between SNOMED CT's conceptual model and ordinary language terms was not only a major bottleneck in processing questions, it presented also a particular challenge for generation where the mapping rules were reversed.

In our prototype system, approach A provides correct answers to 38% of the questions in the corpus. It performs best with simple *What is* and Yes/No questions that can be answered with a simple definition or *Yes* or *No* followed by the reordered input question, and it provided for this type of questions the most reliable answers.

3. Strategy B: Summarization

In our collection of questions on diabetes, we observed that patients often need to know something that cannot be answered by querying the SNOMED CT ontology, no matter how cleverly this is done. Multi-document summarization provides a robust, well-established way to address this problem (Demner-Fushman and Lin, 2006; Lee et al., 2006; Niu et al., 2006). It is a text-to-text generation method that answers questions by finding and manipulating text from documents in a reference corpus. The answer generation happens in a two-step process in which first the most relevant documents (with respect to the question) are found using information retrieval techniques. The top-ranking documents then become input to a summarization component responsible for compiling the final answer. The information retrieval component ranks documents based on the cosine similarity of bag-of-words vectors with tf-idf weighting. For the summarizer we use an implementation of the unsupervised, graph-based LexRank algorithm (Erkan and Radev, 2004) coupled with Maximum Marginal Relevancy (Carbonell and Goldstein, 1998) to ensure information diversity in the answer.

A summary of multiple documents is of course unlikely to be an exact answer to the patient's question. Indeed, for many specific questions, the answer is unlikely to be found at all in the reference collection. In these cases we consider the goal of this approach to be to deliver information pertinent to the question which hopefully allows the patient to infer the answer to his question.

The basic unit is a document in the information retrieval step and a paragraph of text in the subsequent summarization step. This choice implies that the answer cannot be shorter than a paragraph, making it considerably longer than answers obtained from SNOMED CT and question generation (Section 4), which are always a single sentence. Apart from this lower limit, the length of the answer is an adjustable parameter and in the prototype it has been set to a maximum of 100 words. The limit is optimized for questions that solicit advice or ask for explanations as human judges overwhelmingly prefer long answers for these types of questions (Kaisser et al., 2008).

Below we describe how we collected the reference corpus and our strategy for dealing with out-of-vocabulary words in the question.

3.1. Reference corpus

The reference corpus is compiled from various publicly available web sources. It collects the contents of 125 web pages, which have been manually curated and linked to individual questions in the question corpus.

For each document we automatically removed boiler-plate text (e.g., menus and copyright notices) and identified headlines and text content via heuristic rules,[5] resulting in a plain-text file. All downloaded pages were further segmented by headlines so that a section of the text below a headline (and before the next one) would be treated as a separate, more specific, document. Counted this way the total number of documents in the corpus is 556.

From a user's perspective, the reference corpus has a broader coverage of topics (e.g., diet) and provides richer information (e.g., how to check one's blood sugar) than SNOMED CT. Therefore, many questions that fall outside the scope of SNOMED CT have answers in the reference corpus. The reference corpus is well-suited for manner (*how*) and advice questions, because answers may convey useful information that cannot easily be formalized or is not universally true, e.g., practical advice and rules-of-thumb.

3.2. Vocabulary expansion

Although the reference corpus covers a broad range of topics, it does not have a specific answer for each and every question. For instance, many questions are on the topic of food and ask specifically about the feasibility of different dietary choices.

However, *some* of these specific questions have actually been asked before and answered before and so the corpus could easily have an answer for a related question, even though it lacks the answer for the question itself. In this case we say the question is *out of vocabulary* but not *out of topic*. We address the lack of recall due to out-of-vocabulary questions by performing vocabulary expansion at query time.

To motivate this, consider the two related questions in (1) and (2).

(1) Kan jeg spise leverpostej? (Can I eat liver paste?)

(2) Kan jeg spise rullepølse? (Can I eat "rolled meat" sausage?)

[5] https://pypi.python.org/pypi/jusText

Example (3) answers (1) and (2) equally well, but only the topic word of (1) *leverpostej* occurs in the text. Here, expanding the topic word of (2) *rullepølse* to *leverpostej* would allow us to consider the answer for that question as well.

(3) Spis mindre af fede kødprodukter som pølser, bacon, salami, leverpostej og frikadeller. (Eat less of fatty meat products like sausages, bacon, salami, liver paste and rissole).

We retrieve the word expansions by selecting the most similar words in a word embedding space (Mikolov et al., 2013), with a manually tuned threshold for similarity. The embeddings were learned on a corpus consisting of a general language corpus[6] and a specialized diabetes corpus. We obtained the specialized corpus by submitting all questions in our collection as queries to a search engine[7] and for each retrieve all documents in the top 50 matches. Using either of the corpora alone resulted in expansions of a much lower quality than when they were combined.

3.3. Strengths and limitations

The summarization approach as implemented here is fully unsupervised and requires no deep semantic processing, which is an advantage since only limited language resources exist for Danish and performance of state-of-the-art tools for, e.g., parsing and named entity recognition is well below that of comparable tools for English. The use of information retrieval techniques enables the summarization approach to recover relevant answer text from the reference corpus with high recall, and this is further improved by the use of a query expansion component, such as that described above. Unfortunately, the high recall comes at the expense of a lower precision, since the lack of semantic processing may result in, e.g., conflation of homographs, causing too much information to be retrieved. Also, for questions where brief and concise answers are appropriate, the summary will often be too verbose, because the answer length is set to a fixed value for all questions.

In patient question answering the overriding concern is patient safety: Providing an answer which is wrong is much worse than not providing an answer at all. From this perspective the trade-off between recall and precision made by summarization might seem unfortunate. There is, however, a limited degree of error detection built into a summary in that the summary usually provides enough context to allow the user to learn whether the text is in fact addressing the question. This is in contrast to many systems that use sophisticated semantic analysis but where errors in the analysis may go unnoticed by the user.

Further limitations of the summarization approach to QA arise from the need for an up-to-date corpus of verified documents and the fact that summarization, in contrast to knowledge-based approaches, cannot infer new answers on its own. It can, however, as we have seen, provide answers to related questions in case the correct answer cannot be found in the corpus.

[6]Korpus2000: http://ordnet.dk/korpusdk

[7]Microsoft Bing API

In the prototype implementation, the summarization approach provided an answer to 248 out of 321 questions. Of these answers 30% were judged to be correct (covering 24% of the whole question corpus). Compared to the other other strategies, summarization performed best on complex questions involving multiple entitites and relations. However, due to the fixed answer size, the answers were generally too long, mixing correct answer text with peripheral or irrelevant content.

4. Strategy C: Question generation

Question generation (QG) (the task of automatically creating questions from various sources of inputs: texts, databases, etc.) was originally used for educational assessment and intelligent tutoring (cf., e.g., Heilman and Smith, 2010). More recently, it has also been exploited to improve closed-domain QA for languages with scarce resources. Since QG is a relatively new technique in QA, there exist at the moment only very few systems that take advantage of QG (e.g., Bernhard et al., 2012).

The basic idea of the QG approach is to identify sentences of informative documents that can serve as answers to potential questions and transform these sentences into their interrogative forms. All question/answer pairs thus extracted are stored in a database. For question answering, a user's question is identified in the question-answer database, and the corresponding answer is returned.

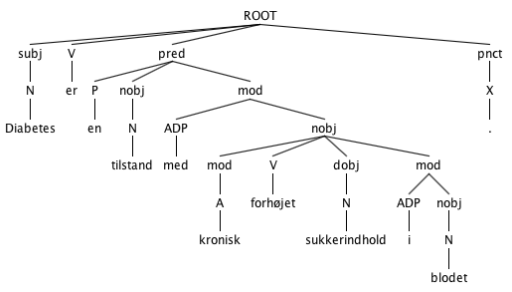
The minimum core resources that are required to produce an operable system are a collection of reliable (authoritative), informative documents, a grammar to parse the documents, and a set of transformation rules that generate questions from the syntactic parse of useful answer sentences. For our prototype we use documents from *Medicinhåndbogen* and *sundhed.dk*, a projective Danish dependency parser (Søgaard and Rishøj, 2010; McDonald and Pereira, 2006), and a set of manually created syntactic transformation rules. To execute the transformation rules, we use Tregex (a tree query language) and Tsurgeon (a tool for modifying trees) (Levy and Andrew, 2006). These tools match the left-hand sides of the transformation rules to syntactic analyses of the documents, apply the syntactic transformations to the document analyses, and output the resulting syntactic descriptions.

As a simple illustration, consider the Danish sentence (4).

(4) Diabetes er en tilstand med kronisk forhøjet sukkerindhold i blodet. (Diabetes is a condition with a chronically elevated level of sugar in the blood.)

and its dependency parse tree in (5).

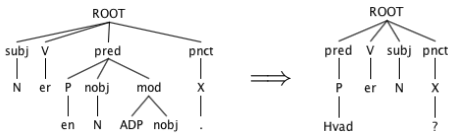
(5)



This sentence provides a useful answer to the question

Hvad er diabetes? (What is diabetes?). The dependency tree for the question is then produced by running the Tregex and Tsurgeon scripts for the transformation rule in (6).

(6)



4.1. Refinements and optimizations

There are a number of refinements and optimizations to the basic setup that can improve the overall performance of the QG module. However, only some of them have been implemented (to some extent). This is because the required resources and tools were either not available for Danish or not adaptable within the project funding and timeframe.

Automatic recognition of potential question topic terms and question foci The term *question topic* is usually used to refer to the object or event someone intends to increase his/her knowledge about by using the question (like *diabetes*, *hypoglycemia*, etc.), while the *question focus* is the term/phrase indicating the semantic type of which the answer is an instance (e.g., *country* in *Which country is most densely populated?*). In order to identify potentially useful answer sentences in a collection of documents, it is therefore extremely beneficial to tag all topic terms and to annotate all named entities with their focus category. This can be accomplished with a named entity recognizer that supports the required classification (cf., e.g., Prager et al., 2006). However, since such an advanced named entity recognizer is not available for Danish, we relied for the prototype on an existing list of on-topic terms (from the Steno Diabetes Center) and a manually created list of focus terms/classes (like *disease*, *condition*, etc.).

Paraphrase generation In several QA systems paraphrasing has been used to better cope with the linguistic variability of questions and answers. Paraphrasing increases the likelihood of finding an answer if the user's question is meaning equivalent to a question in the database but varies from it in linguistic form. Because of time constraints, we focused first on simple paraphrases (taken from Microsoft Word's thesaurus) that could safely be produced by global substitution without much regard for the surrounding context. However, for the generation of more complicated context-sensitive and phrasal paraphrases, several promising approaches extract paraphrases automatically from the Web (cf., e.g., Duclaye et al., 2003) or exploit phrase tables from existing machine translation systems (cf., e.g., Kuhn et al., 2010).

Answer selection To further improve recall, a very rudimentary fuzzy matching procedure has been incorporated. This, however, leaves room for improvements. A lemmatizer, for example, can help in identifying the answer in cases where a user's question differs only morphologically from a question in the database, as in *Hvad er symptomer på diabetes?* and *Hvad er symptomerne på diabetes?* (What are (the) symptoms of diabetes?). There are, of course, many other algorithms that can be used to compute the best matching database question, ranging from relatively simple algorithms that compute the edit distance between two strings (e.g., Levenshtein distance computing algorithm) to more sophisticated syntactic similarity measures (cf., e.g., Croce et al., 2011).

Often there will be more than one answer to a question. Indeed, if the document collection is rich enough, there may easily be ten or twenty options for answering common queries such as *What is diabetes?* Even though we did not rank the answers, there are a number of algorithms for selecting the "best" answer, among them ranking algorithms that exploit syntactic and semantic features to account for both linguistic quality and information content at the same time (cf., e.g., Athenikos and Han, 2010).

Co-reference chain detection If sentences following an identified answer sentence further elaborate that sentence, these sentences are often linked through a co-reference chain. Consider, for instance, the following English translation of an extract from a document:

(7) In the pancreas *hormones*$_1$ are produced, which control, among others, the body's sugar balance. *These*$_1$ are glucagon and *insulin*$_2$. The *insulin*$_2$ enables sugar in the blood to enter more easily, among others, muscle and liver cells.

Here, QG produces from the first sentence the question *What is produced in the pancreas?* Co-reference chain detection can then be used to identify the whole extract as a more detailed answer than the one provided by only the first sentence. Unfortunately, there are no practically applicable co-reference resolution systems specifically designed for Danish; there are a few language independent systems (cf., e.g., Zhekova and Kübler, 2010; Zhekova et al., 2012) that have been demonstrated to work effectively for several different languages. However, because of the scope of the project, it was not possible to adapt the language dependent modules to the given language data.

Ordering for multiple sentence answers If there are many options for answering common queries such as *What is diabetes?* often several of them can be combined to a more comprehensive answer. Consider, for example, the translations of two Danish sentences that generate *What is diabetes?* (in Danish).

(8) Diabetes is a condition with chronically high level of sugar in the blood. Diabetes is a serious disease because it causes a significant risk for developing complications of the eyes, kidneys, nerves, and blood vessels.

In this case a combination is desirable. However, nothing prevents two sentences mapping to the same question from being uninformative variants of each other. To ensure that new information is actually added, techniques from automatic summarization can be used. To maximize the information content of combined answers, we can, for instance, select the answer that, for a fixed size, has the largest possible number of different concepts (e.g., bigrams) (Gillick and Favre, 2009).

Figure 2: Screenshot of the system's output for the question *Hvad er hypoglykæmi?* (What is hypoglycemia?).

4.2. Strengths and limitations

Even without most of the refinements mentioned above the QG approach is able to produce a variety of more complex interrogative questions, including causative questions, that are assumed to be inherently more difficult than other interrogative questions, like, for example, factoid and definitional questions. English translations of a few typical examples are given in (9a–f).

(9) a. *What are the symptoms of diabetes?* The classic symptoms of untreated diabetes are weight loss, increased urination (polyuria), and increased sense of thirst (polydipsia) and hunger (polyphagia).

b. *Is diabetes contagious?* Diabetes is not contagious.

c. *Can diabetes be cured?* Diabetes cannot be cured, but there are now drugs that can increase insulin production and increase the cells' sensitivity to insulin.

d. *What is the diabetes treatment aiming at?* The diabetes treatment aims at bringing the level of blood glucose as near as possible to normal to eliminate the symptoms of high blood glucose and to prevent the development of complications.

e. *What causes diabetes?* Diabetes is caused by defective insulin secretion, reduced insulin action or a combination of these factors.

f. *When is a person diagnosed with diabetes?* A person is diagnosed with diabetes when the glucose levels are not normally controlled and the concentration is too high.

However, the scope of QG is limited to encyclopedic one sentence questions. In some cases it is possible to produce multi-clausal questions such as *What happens if the blood sugar is too low?* but these typically make up only a small portion of the generated questions.

We created altogether 45 transformation rules (the number could have been slightly reduced by fully exploiting the notational devices that Tregex and Tsurgeon provide). These generated from a small document collection (altogether 750 sentences) a total of 148 questions. Surprisingly, only 16% of them were contained in the question corpus (7.5% of the entire corpus), although all generated questions are perfectly reasonable to ask. Thus for real-life health-related questions posed by lay people (Kilicoglu et al., 2013), QG seems to be more like a back-up strategy when other strategies, for whatever reason, fail to provide meaningful answers to Wikipedia-related questions.

5. Evaluation

The three approaches are embedded in a prototype system which runs them in parallel and provides results for all of them. Figure 2 shows a screenshot of the system's output for the question *Hvad er hypoglykæmi?* (What is hypoglycemia?). The output shows (i) the suggested answer from each approach, (ii) the confidence scores, and (iii) the comparable scores in percentages. In this example approach C provides the answer that is given to the user (because it received the highest score). The English translations are as follows (less relevant information provided by approach B is displayed in gray):

(A) Hypoglycemia is a type of hypoglycemic disease and is physically related to the endocrine structure,

(B) It is essential whether you have had severe hypoglycemia. Hypoglycemia is assessed as severe hypoglycemia if you have needed assistance. A person who experiences more than one case of severe hypoglycemia within a 12-month period, does not meet the requirements for the issuance, renewal, and maintenance of a driver's license. [420-1] Hypoglycemia or low blood sugar in a diabetes patient is an acute condition where the level of blood sugar (glucose) is lower than normally observed in people without diabetes. Hypoglycemia is usually caused by diabetes treatment (insulin or tablets), and hypoglycemia is usually accompanied by a variety of symptoms. By consuming glucose, symptoms usually disappear within 10–15 minutes. Hypoglycemia can also occur in non-diabetics. [436-1] You need less insulin: [436-7] What can you do yourself? [436-14] What is hypoglycemia and what are the symptoms of hypoglycemia? [38-0],

(C) Hypoglycemia, or hypo as most diabetics call it, is a physical reaction when the blood sugar drops to a level below 2–5 mmol/l.

The approaches were evaluated on our corpus of 321 questions. 306 questions got an answer from at least one approach, and at least one of the approaches answered correctly in 45% of these cases (43% of the total corpus). This quality is too low for a running system. However, as Danish is a less-resourced language with much fewer medical knowledge resources than English, the prototype is a step toward the implementation of a medical QA system for the Danish citizens.

6. Consequences and observations

The goal of our strategy A (Section 2) was to explore to what degree it is possible to build a natural language interface to the SNOMED CT ontology. This required a mapping between layman's everyday language and the terminology of professionals. However, this mapping is not first and foremost a technical challenge, but rather a conceptual challenge, because professionals and patients understand and conceptualize medicine in fundamentally different ways.[8] Because of the divergent conceptualizations of the world and SNOMED CT's relational sparseness, it has been an extremely difficult task to disambiguate and interpret laymen's questions in the SNOMED CT ontology. Moreover, a considerable amount of user questions could not be processed since they simply could not be adequately interpreted in the lightweight description logic SNOMED CT is based on.

The problem of differences in terminology is less severe in the text-to-text generation strategies B (summarization) and C (question generation), for two reasons. First, answers are located by matching words in the question with words in the documents, ensuring a common vocabulary. Second, the documents in our collection are written with a layman audience in mind, addressing the concerns of patients and not professionals. But this issue does also affect these systems and may become more pressing when we move beyond small curated document collections and use text written for a broader range of audiences, including health professionals. However, even based on our rather small document collections, these approaches revealed serious limitations: summarization suffers from precision issues and question generation from recall issues.

Moreover, by considering the questions of our corpus of real-life layman's questions that our strategies were not able to adequately deal with, it became quite obvious that their complexity is way beyond currently known QA technology.

[8] A point also raised by Udo Hahn during a panel discussion at Medinfo 2013.

7. Acknowledgments

The ESICT project, running from 2010 – 2013, was funded by the Danish Strategic Research Council. We would like to thank our project partners: Ulrich Andersen, Anna Braasch, Søren Brunak, Anders Jensen, Lars Kayser, Ole Norgaard, and Stefan Schulz for contributing important information to this paper.

8. References

Andersen, U., Braasch, A., Henriksen, L., Huszka, C., Johannsen, A., Kayser, L., Maegaard, B., Norgaard, O., Schulz, S., and Wedekind, J. (2012). Creation and use of language resources in a question-answering eHealth system. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2536–2542, Istanbul.

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24.

Bernhard, D., de Viron, L., Moriceau, V., and Tannier, X. (2012). Question generation for French: Collating parsers and paraphrasing questions. *Dialogue and Discourse*, 3(2):43–74.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne.

Croce, D., Moschitti, A., and Basili, R. (2011). Semantic convolution kernels over dependency trees: Smoothed partial tree kernel. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2013–2016, New York, NY.

Demner-Fushman, D. and Lin, J. (2006). Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual meeting of the Association for Computational Linguistics*, pages 841–848, Sydney.

Duclaye, F., Yvon, F., and Collin, O. (2003). Learning paraphrases to improve a question-answering system. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop in NLP for QA*, pages 35–41, Budapest.

Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Gillick, D. and Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, CO.

Guo, Y., Wang, H., and van Genabith, J. (2011). Dependency-based n-gram models for general purpose sentence realisation. *Natural Language Engineering*, 17(4):455–483.

Heilman, M. and Smith, N. A. (2010). Good question? Statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings*, pages 609–617, Los Angeles, CA.

Jongejan, B. and Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153, Suntec.

Kaisser, M., Hearst, M., and Lowe, J. (2008). Improving search results quality by customizing summary lengths. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 701–709, Columbus, OH.

Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2013). Interpreting consumer health questions: The role of anaphora and ellipsis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 54–62, Sofia.

Kuhn, R., Chen, B., Foster, G., and Stratford, E. (2010). Phrase clustering for smoothing TM probabilities – or, how to extract paraphrases from phrase tables. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 608–616, Beijing.

Lee, M., Cimino, J., Zhu, H., Sable, C., Shanker, V., Ely, J., and Yu, H. (2006). Beyond information retrieval – Medical question answering. *AMIA Annual Symposium Proceedings*, 2006:469–473.

Levy, R. and Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2231–2234, Genoa.

McDonald, R. and Pereira, F. (2006). Online learning of approximate dependency parsing algorithms. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–88, Trento.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv e-prints:1301.3781*, pages 1–12.

Niu, Y., Zhu, X., and Hirst, G. (2006). Using outcome polarity in sentence extraction for medical question-answering. *AMIA Annual Symposium Proceedings*, 2006:599–603.

Prager, J. M., Chu-Carroll, J., Brown, E. W., and Czuba, K. (2006). Question answering by predictive annotation. In Strzalkowski, T. and Harabagiu, S., editors, *Advances in Open Domain Question Answering*, pages 307–347. Springer, Dordrecht.

Søgaard, A. and Rishøj, C. (2010). Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1065–1073, Beijing.

Zhang, Y. (2010). Contextualizing consumer health information searching: An analysis of questions in a social Q & A community. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 210–219, Arlington, VA.

Zhekova, D. and Kübler, S. (2010). A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala.

Exploring Negation Annotations in the DrugDDI Corpus

Behrouz Bokharaeian*, Alberto Diaz*, Mariana Neves†, Virginia Francisco*
*NIL Group, Universidad Complutense de Madrid
Madrid, Spain
behroubo@ucm.es, albertodiaz@fdi.ucm.es, virginia@fdi.ucm.es
†Hasso-Plattner-Institute at the University of Potsdam, Germany
marianalaraneves@gmail.com

Abstract

Detecting drug-drug interactions (DDI) is an important research field in pharmacology and medicine and several publications report every year the negative effect of combining drugs and chemical treatments. The DrugDDI corpus is a collection of documents derived from the DrugBank database and contains manual annotations for interactions between drugs. We have investigated the negated statements in this corpus and found that they consist of approximately 21% of its sentences. Previous works have shown that considering features related to negation can improve results for the DDI task. The main goal of this paper is to describe the process for annotating the DDI-DrugBank corpus with negation cues and scopes, to show the correlations between these and the DDI annotations and to demonstrate that negations can be used as features for a DDI detection system. Basic experiments have been carried out to show the benefits when considering negations in the DDI task. We believe that the extended corpus can be a significant progress in training and testing algorithms for DDi extraction.

Keywords: Negation detection, Drug-Drug Interaction, Relation extraction

1. Introduction

A drug-drug interaction (DDI) usually occurs when one drug changes the level of activity of another drug. According to FDA's reports and acknowledged surveys (Gurwitz et al., 2000), over 2 million serious Adverse Drug Reactions (ADRs) occur in the United States every year, including the register of one hundred thousand deceases (Lazarou et al., 1998). Moreover, 3.5% of these deaths are due to drug-drug interaction (Martin, 1990). Detecting and identifying interactions between drugs is a crucial field of research given the high risks of most drug-drug interactions and the importance of patient safety and health care cost control. Many academic researchers and pharmaceutical companies have developed databases where DDI are recorded, but most of the research and valuable information is still only found in unstructured text documents, such as scientific publications and technical reports.

Information extraction is an important task in natural language processing (NLP) and has also been used in many applications in the biomedical domain, ranging from simple binary relationships to complex and hierarchical relation extraction (McDonald et al., 2005). Recent research on biomedical information extraction has focused on biological entities and relationships, since many annotated corpora are available for this purpose, which are valuable resources for repeatable automatic training and evaluation of NLP techniques. For instance, several corpora have been annotated for protein-protein or gene-protein interactions, such as Aimed (Bunescu et al., 2005), LLL (Nedellec, 2005), IEPA (Ding et al., 2002) or BioCreAtIvE-PPI(Krallinger et al., 2008)).

A DrugDDI corpus was initially developed by (Segura-Bedmar et al., 2011a) based on a set of 579 xml files describing DDIs which was randomly collected from the DrugBank database (Wishart et al., 2007). The UMLS MetaMap (MMTx) tool (Aronson, 2001) was used to analyse the corpus and was manually annotated with the help of pharmacist experts (DDI 2011 corpus). With the aim of encouraging researchers to explore new methods for extracting drug-drug interactions, the first DDI Extraction challenge task[1] was held in 2011 with the participation of ten teams (Segura-Bedmar et al., 2011b). The best results were an F-measure of 65.74%, a precision of 65.04% and a recall of 71.92% in detecting and classifying DDIs (Thomas et al., 2011). A second challenge was held on 2013 as part of SemEval: the DDI Extraction2013[2]. A new corpus was developed which included the corpus used in 2011 (DDI-DrugBank 2013) as well as Medline abstracts. Participating teams developed solutions based on supervised and sentence-level relation extraction methods and the best F1 score obtained was 80%. According to Segura and her colleagues, increasing the size of the corpus and optimizing the quality of annotations have contributed to this improvement (Segura-Bedmar et al., 2013).

The DDI-DrugBank 2013 corpus is a useful resource for performing comparable experiments and for investigating relation extraction methods. However, one limitation of this corpus is the lack of negation annotation. For instance, in the sentence below, an interaction between *itraconazole* and *S-ketamine* drugs could be identified if negation is ignored.

Ticlopidine treatment increased the mean area under the plasma concentration-time curve extrapolated to infinity (AUC(0-)) of oral ketamine by 2.4-fold, whereas itraconazole treatment did not increase the exposure to S-ketamine.

Negation is frequently used in clinical and biomedical documents and it is an important cause of low precision in automated indexing systems (Chapman et al., 2002). For instance, Chapman has observed that 95% to 99% of the searched reports would state *no signs of fracture* or similar

[1] http://labda.inf.uc3m.es/DDIExtraction2011/

[2] http://www.cs.york.ac.uk/semeval-2013/task9/

expressions in a certain radiology report database (Chapman et al., 2002). As a result, identifying negative statements is an important task to obtain accurate knowledge from textual data.

In previous work (Bokharaeian et al., 2013), the DDI 2011 corpus was annotated with negations (NegDrugDDI corpus) and a basic experiment showed that improvements in drug-drug interactions can be obtained when considering annotations for negations. Additionally, the best-scoring team in the DrugDDI 2013 challenge also use negation information in their system (Chowdhury et al., 2013).

In this paper, we describe the annotation of the DDI-DrugBank 2013 corpus with negation cues and scopes, the hereafter called NegDDI-DrugBank 2013, following the BioScope guidelines (Szarvas et al., 2008). We also present the correlations between the negation annotations and the position of the drugs in a sentence. Finally, we have performed some experiments with the TEES event extraction tool (Bjorne and Salakoski, 2013) to confirm the positive effect of the negation annotations for the DDI task.

In Section 2, we present related work on previous corpora annotated with negation, while Section 3 describes corpora annotated with drug-drug interactions. In Section 4, the annotation process and the obtained results are described. Section 5 presents the correlations between DDI and negation that have been found in the extended corpus while Section 6 shows the experiments carried out to confirm the effects of negation annotations for the DDI task. Finally, Section 7 presents discussions and suggestions for future works.

2. Corpora annotated with negation

In this section, we review the main corpora annotated with negation, emphasizing the annotation guidelines that were followed and the main differences between them.

2.1. Bioscope

Bioscope[3] (Szarvas et al., 2008) is an open access corpus of biomedical documents, manually annotated with negation and speculation. It contains more than 20,000 sentences which are split in three collections: clinical documents (6,383 sentences, 863 with negations), scientific papers (2,670 sentences, 339 with negations) and scientific abstracts (11,871 sentences, 1,597 with negations). All the sentences which assert the non-existence of something are annotated, including sentences which do not contain any biomecial term. Each negated sentence is annotated with information about the negation cue and the scope of negation.

The annotation of Bioscope followed a min-max strategy: the minimal unit that expresses negation is considered the negation cue (min strategy) and the scope is extended to the largest syntactic unit possible (max strategy). The negation cue is always included in the scope. However, it is worth emphasizing that when the scope is opened at the cue and continues to the right of the cue (around 90% of the cases), the scope affected by the cue leaves the subject out. This corresponds to sentences in active voice which are the most frequent case. Additionally, there are cases in which the scope is opened to the left of the cue. The most frequent ones are the structures in passive voice. As shown in (Szarvas et al., 2008), passive voice is an exception in the way of tagging sentences in Bioscope. In this case, the subject is annotated within the scope, because if the sentence had been written in active voice, it would have been the object of a transitive verb.

[3] http://www.inf.u-szeged.hu/rgai/bioscope

2.2. SFU Review Corpus

SFU Review Corpus (Konstantinova et al., 2012) is a freely available corpus annotated with negation and speculation. It consists of 400 documents of movie, book and consumer product reviews. It is annotated with negative and speculative keywords and their scope. The entire corpus was manually annotated by one linguist and reviewed by another one. The guidelines followed during the annotation was an adaptation of Bioscope guidelines, which main changes were:

- Negation cues were not included in the scope.
- Coordination was annotated in a different way.
- Sentences with negation cue and without scope were possible.

2.3. ConanDoyle-neg

ConanDoyle-neg (Morante and Daelemans, 2012) was released in conjunction with the 2012 shared task on NR hosted by The First Joint Conference on Lexical and Computational Semantics (*SEM 2012). It is a corpus of Arthur Conan Doyle's stories manually annotated with negation cues and their scope. The annotation was performed by two annotators using the Salto Tool.

The following is annotated in each sentence which contain negation statements: the negation cue, its scope and the negated event. For example, in the sentence *"After mine I asked no questions" no* is identified as the negation cue, *after mine I asked questions* is identified as the scope and *asked* is the negated event.

This corpus annotation was inspired by the guidelines of Bioscope, but with several differences, being the following the most important ones:

- The negated event is annotated.
- Negation cues are not included in the scope.
- Scopes can be discontinuous.
- All arguments of the negated event are included in the scope, including the subject (which in Bioscope corpus was kept out in active sentences).
- Affixal cues are annotated. If the scope of a negation cue is not explicit, the negation cue is marked as such, but the scope is not annotated. If the scope is recoverable from the same sentence, it is added to the scope.

The domain of this corpus is very restrictive so some constructions that are typical in other domains are left out. For instance, constructions that express absence of an entity, which are very frequent in biomedial texts, are not included in this corpus.

2.4. UAM Spanish Treebank

UAM Spanish Treebank (Moreno Sandoval et al., 2003) is a corpus composed of 1,501 syntactically annotated sentences derived from Spanish newspapers. The syntactic annotation was extended with annotations for negation. Annotation of negation was carried out by two experts in Corpus Linguistics. The annotation guidelines were very similar to those of Bioscope, except for one main difference: all arguments of the negated events are included in the scope, including the subject (which were kept out in active sentences in the Bioscope corpus).

3. Drug-drug interaction annotation

The DDI 2011 corpus was the first annotated corpus dealing with the interaction phenomenon between drugs. The corpus was designed by (Segura-Bedmar et al., 2011a) in order to encourage the NLP community to conduct further research in the field of pharmacology. A set of 579 xml files describing DDIs was randomly collected from the DrugBank database (Wishart et al., 2007). The corpus was analyzed by the UMLS MetaMap tool (MMTx) (Aronson, 2001) and was manually annotated with the help of pharmacist experts.

This corpus is provided in the unified format used for PPI corpora proposed in (Pyysalo et al., 2008) (see Figure 1). Each entity (drug) includes reference (origId) to the id phrase in the MMTX format corpus text in which the corresponding drug appears. For each sentence in the corpus, all DDI candidate pairs are generated from the possible combination of different drugs appearing therein. Each DDI candidate pair is represented as a *pair* node in which the ids of the interacting drugs are registered in its *e1* and *e2* attributes. If the pair is a DDI, the *interaction* attribute must be set to *true*, otherwise this attribute must be set to *false*.

The DDI-DrugBank 2013 corpus was developed for the DDI Extraction 2013 SemEval task and includes part of the the DDI 2011 corpus. Concretely, new documents were annotated from the DrugBank database and were used for the test dataset (DDI-DrugBank Test 2013 corpus), while 572 documents from the previous corpus were used as training dataset (DDI-DrugBank Train 2013 corpus). Therefore, the DDI-DrugBank 2013 corpus contains a total of 730 documents. A dataset of 233 MedLine abstracts (DDI-MedLine 2013 corpus) was also annotated for the 2013 shared task, however, in this work we have concentrated on the DrugBank documents.

Table 1 shows basic statistics of the DDI-DrugBank 2013 corpus. It contains 6,648 sentences with 9.1 sentences per document on average. The average number of drug mentions per document was 21.15, and the average number of drug mentions per sentence was 2.4. Finally, among the 31,270 candidate drug pairs, only 4,672 (14.94%) were annotated as positive interactions, (i.e., DDIs), while 26,598 (85.06%) were marked as negative interactions (i.e., non-DDIs). There is a much larger proportion of negative instances than positive ones.

All drug-drug interactions in the DDI-DrugBank 2013 corpus was also annotated with one of the following four interaction types: *advice*, *effect*, *mechanism* and *int*. The *advice* type corresponds to an advice or recommendation regarding the concomitant use of the two drugs, the *effect* category refers to the effect of DDIs, the *mechanism* type were assigned to DDIs which describe pharmacodynamic or pharmacokinetic mechanism and the default *int* category is used otherwise. More detailed definition of the types can be found at (Segura-Bedmar et al., 2013). With respect to the distribution of categories, as can be seen in Table 2, there is a smaller number of instances for categories *int* and *advice* and *effect* type is the most frequent.

4. Annotating DDI-DrugBank corpus with negation

The aim of this paper is to extend a drug-drug interactions corpus (DDI-DrugBank 2013) with annotations for negation, the NegDDI-DrugBank 2013, as none of the existing corpora meets this requirement. All the sentences in the original corpus were annotated, which conforms 6,648 sentences from 730 files. For the DDI DrugBank 2013 training dataset, annotations from the NegDrugDDI corpus (Bokharaeian et al., 2013) have been transferred to the NegDDI-DrugBank 2013 corpus and then reviewed, given that there were some discrepancies between the documents from the two DDI editions.

For the DDI DrugBank 2013 test dataset, a first annotation was done with the rule based system (Ballesteros et al., 2012), which follows the BioScope guidelines to annotate sentences with negation. The annotation consisted on adding two new tags, the cue and the scope of the negations, as depicted in Figure 3. The pre-annotation automatically obtained was then reviewed by four annotators using the Brat NLP annotation tool[4]. Brat is a web based software tool which was developed for rich annotating which has proven to decrease the annotation time and to increase the quality of the resulting annotations (Stenetorp et al., 2012). A screenshot of the NegDDI-DrugBank 2013 corpus as visualized in the tool is shown in Figure 2. The test dataset was split in four parts, one for each annotator, who have manually corrected the automatically generated annotations, whenever necessary, and have added the missing ones. Subsequently, the more experienced annotator reviewed all the annotations to ensure coherence. According to the annotators, 18 modifications have been done. That is, the algorithm have annotated the majority of the sentences correctly. The extended corpus is available for public use [5].

We have performed an analysis on the number of distinct cues in the entire NegDDI-DrugBank 2013 corpus and the number of different problematic annotation that were observed. This analysis is shown in Table 3. As can be seen in this table, *not* and *no* are by far the most frequent cues in the corpora: 1018 and 498 occurrences. However, more changes have been performed with cue *not*, 27.41% of changes. On the other hand, it can be observed that the most problematic cue is *neither ... nor ...*, with a 85.71% of changes. It is due to the difficult double cue pattern associated to this cue. Most of the errors with the other cues are associated with problems detecting certain patterns of pas-

[4] http://brat.nlplab.org/

[5] http://nil.fdi.ucm.es/sites/default/files/NegDDI_DrugBank.zip

```
−<sentence id="DrugDDI.d346.s0" origId="s0" text="Uricosuric Agents: Aspirin may decrease the effects of probenecid,
sulfinpyrazone, and phenylbutazone.">
  <entity id="DrugDDI.d346.s0.e0" origId="s0.p0" charOffset="0-17" type="drug" text="Uricosuric Agents"/>
  <entity id="DrugDDI.d346.s0.e1" origId="s0.p2" charOffset="19-26" type="drug" text="Aspirin"/>
  <entity id="DrugDDI.d346.s0.e2" origId="s0.p6" charOffset="55-65" type="drug" text="probenecid"/>
  <entity id="DrugDDI.d346.s0.e3" origId="s0.p7" charOffset="67-81" type="drug" text="sulfinpyrazone"/>
  <entity id="DrugDDI.d346.s0.e4" origId="s0.p9" charOffset="87-101" type="drug" text="phenylbutazone"/>
  <pair id="DrugDDI.d346.s0.p0" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e1" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p1" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e2" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p2" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e3" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p3" e1="DrugDDI.d346.s0.e0" e2="DrugDDI.d346.s0.e4" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p4" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e2" interaction="true"/>
  <pair id="DrugDDI.d346.s0.p5" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e3" interaction="true"/>
  <pair id="DrugDDI.d346.s0.p6" e1="DrugDDI.d346.s0.e1" e2="DrugDDI.d346.s0.e4" interaction="true"/>
  <pair id="DrugDDI.d346.s0.p7" e1="DrugDDI.d346.s0.e2" e2="DrugDDI.d346.s0.e3" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p8" e1="DrugDDI.d346.s0.e2" e2="DrugDDI.d346.s0.e4" interaction="false"/>
  <pair id="DrugDDI.d346.s0.p9" e1="DrugDDI.d346.s0.e3" e2="DrugDDI.d346.s0.e4" interaction="false"/>
</sentence>
```

Figure 1: The unified XML format of a sentence in the DrugBank-DDI 2013 corpus.

	Number	**Avg. per document**
Documents	730	
Sentences	6,648	9.11
Entities	15,441	21.15
Candidate drug pairs	31,270	42,84 (4.70 per sentence)
Positive interactions (DDIs)	4,672	6.40 (14.94%)
Negative interactions (no DDIs)	26,598	36.44 (85.06%)

Table 1: Basic statistics of the DDI-DrugBank 2013 corpus.

Training	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	26005	22217	3788	1535	1257	818	178
Test	pairs	negative DDIs	positive DDIs	effect	mechanism	advice	int
DrugBank	5265	4381	884	298	278	214	94

Table 2: Statistics of the training and test datasets of the DDI-DrugBank 2013 corpus.

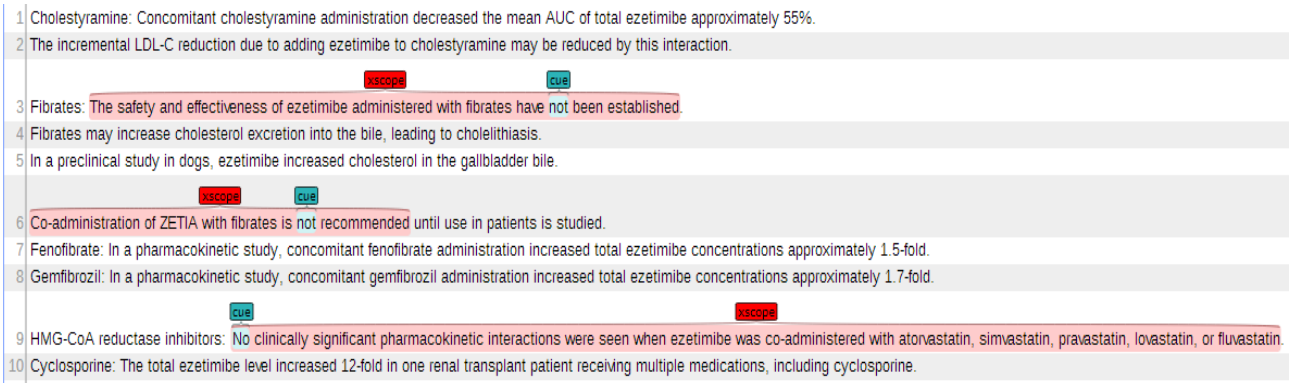


Figure 2: Examples of negation cue and scope annotations.

Cue	**DDI-DrugBank Train**	**Changes**	**DDI-DrugBank Test**	**Changes**	**DDI-DrugBank**	**Total**	**Rate**
not	855	266	163	13	1018	279	27.41%
no	439	58	59	1	498	59	11.85%
without	47	8	9	4	56	12	21.43%
neither ... nor ...	14	12	0	0	14	12	85.71%
absence	10	5	3	0	13	5	38.46%
lack	8	1	0	0	8	1	12.50%
cannot	7	4	3	0	10	4	40.00%
Total	1380	354	237	18	1617	372	23.01%

Table 3: Statistics of the negative cues in the training and test datasets, the changes for each cue during manual checking and the rate of changes, for the NegDDI-DrugBank 2013.

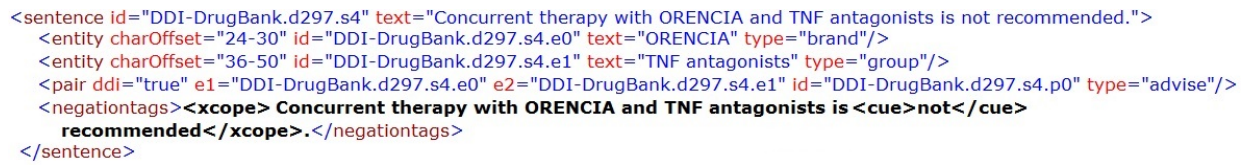


Figure 3: The extended unified XML format of a sentence with negation cue in NegDDI-DrugBank corpus.

sive voice sentences and with the bad processing of commas and copulative keywords.

5. Analysis of correlations between negations and DDI annotations

NegDDI-DrugBank 2013 corpus contains 1,448 sentences with at least one negation scope, which correspond to 21.78% of the sentences (4). This confirm the statement that negation is frequently used in clinical and biomedical documents, and particularly, in pharmacological documents describing drug activity.

Table 5 shows the correlations between the annotations for negation scopes and the position of the two candidate drugs that represent a DDI. The first two columns indicate the position of the drugs, there are 5 possibilities:

- both drugs inside of the negation scope (inside, inside).
- both drugs outside of the negation scope but on the right hand side of the sentence (right, right).
- both drugs outside of the negation scope but on the left hand side of the sentence (left, left).
- one drug inside the negation scope but the other one outside on the right-hand side (inside, right).
- one drug inside the negation scope but the other one outside on the left hand side (inside, left).

For instance, Figure 3 shows a sentence with a negation cue and two drug which are both inside of the negation scope. We can conclude from this data that, in the majority of the cases (around 90%), there is no DDI when a negation scope is present. With respect to the position of the drugs, the best correlation occurs when both drugs are inside the negation scope (93.78%), while the worst correlation occurs when one drug is inside and the other one is outside or on the right hand side (88.43%).

The correlation between DDI type and drug positions compared to negation scope has also been analyzed. As Table 6 confirms, there is a clear correlation between the DDI type and relative candidate drug positions to negation scope. The highest correlation can be seen when both candidate drugs are inside the negation scope and DDI type is *advice* (78.65% of all *advice* type cases with negation cue mention a positive DDI). For instance in the below sentence:

<xcope>It is recommended that the combination of intravenous ***dantrolene sodium*** *and calcium channel blockers, such as* ***verapamil****, <cue>not </cue>be used together during the management of malignant hyperthermia crisis until the relevance of these findings to humans is established.</xcope>*

The candidate drugs (*dantrolene sodium* and *verapamil*) are both inside of the negation scope and the *advice* type was assigned to the DDI.

The other three DDI types (*effect*, *mechanism* and *int*) have a similar behavior regarding the correlation between DDI type and candidate drug positions. For instance, for all of them, percentages are low (effect= 4.49%, mechanism=16.8% and 0% for int) when two candidate drugs are inside the scope.

Table 7 shows the average of correlations between the DDI type and candidate drug positions. As can be seen there is a significant difference between *advice* type and the other three DDI types. The 53.87% of the sentences with negation that contains a positive DDI correspond to *advice* type and the 1.3% of the sentences with negation that contains a positive DDI correspond to *int* type.

We can conclude that the position of entities regarding the scope of negation is an important factor in determining the effect of negation and the candidate DDIs.

On the other hand, regarding to Drug-Drug Interaction relation, *recommended* and *advised* words have negative polarity. In fact *recommendation* is used to avoid coadministration of two drugs, instead of recommending them, but *effect*, *excretion* and *interact* phrases have positive polarities. Consequently, classifying and extracting positive DDIs should consider these important factors in addition to other syntactic factors that are usually employed.

Our analysis shows that we need semantic and polarity-based processing to efficiently employ negation information in relation extraction task. For instance, the two sentences below are in passive voice and they have similar length and annotations for negations. The first one mentions a drug-drug interaction in an advisory notion, while the second one explains a mechanism for possible drug interaction, but does not mention a DDI. In both sentences, two drug names are inside the negation scope and related verb and adverbs are also inside of scope. These two sentences are good examples that deep and semantic processing are needed to employ negation in detecting positive drug-drug interactions.

- <xcope>Concurrent therapy with ORENCIA and TNF antagonists is <cue>not</cue> recommended.</xcope>
- <xcope>This small decrease in ec of gabapentin by cimetidine is <cue>not</cue> expected </xcope> to be of clinical importance.

	Number	Percentage (%)
Documents	730	
Sentences	6,648	
Sentences with negation	1,448	21.78
Sentences without negation	5200	78.22

Table 4: Basic statistics from the NegDDI-DrugBank 2013 corpus

Drug1position	Drug2position	DDI	Train	Test	Total	Percentage (%)
inside	inside	false	613	730	1343	93.78
inside	inside	true	39	50	89	6.63
left	left	false	141	1191	1332	89.82
left	left	true	27	124	151	11.34
right	right	false	101	819	920	92.56
right	right	true	12	62	74	8.04
inside	left	false	256	921	1177	92.31
inside	left	true	6	92	98	8.33
inside	right	false	52	437	489	88.43
inside	right	true	7	57	64	13.09

Table 5: Correlations between DDI and drug positions compared to negation scope for the NegDDI-DrugBank 2013 corpus. The third column indicates if the candidate DDI associated with the annotation is true or false. The fourth and fifth columns indicate if the correlation appears in the training or in the test dataset. Finally, the last column indicates the total of possible correlations of each type and the corresponding percentage.

Drug1position	Drug2position	Type	Total	Percentage (%)
inside	inside	advise	70	78.65
left	left	advise	50	33.11
right	right	advise	24	32.43
inside	right	advise	37	57.81
inside	left	advise	66	67.34
inside	inside	effect	4	4.49
left	left	effect	56	37.08
right	right	effect	14	18.91
inside	left	effect	26	26.53
inside	right	effect	15	23.43
inside	inside	mechanism	15	16.85
left	left	mechanism	44	29.13
right	right	mechansim	34	45.94
inside	left	mechanism	6	6.12
inside	right	mechanism	10	15.62
inside	inside	int	0	0
left	left	int	1	0.66
right	right	int	2	2.7
inside	left	int	0	0
inside	right	int	2	3.12

Table 6: Correlations between positive DDI and drug position compared to negation scope. The last columns show the total of possible correlations of each type and the corresponding percentage.

Type	Total	Average (%)
advise	247	53.87
effect	115	22.01
mechanism	129	26.81
int	5	1.3

Table 7: Total average correlations between DDI type and candidate drug positions

6. Exploring negation features

In addition to extending the DDI-DrugBank 2013 corpus, we carried out experiments using the version 2.1 of TEES event extraction software tool[6] to verify the effects of the negation annotations in a relation extraction task. TEES is a well known machine-learning based tool for extract-

[6] http://jbjorne.github.io/TEES/

ing text-bound graphs from natural language text and has shown successful performance in many binary relationship and event extraction tasks (Bjorne and Salakoski, 2013).
TEES supports negation detection using the schema used in the BioNLP Genia event Extraction tasks[7], where a negation attribute is assigned to the event, but no cue or scope are annotated. When performing our experiments with TEES, we have added the negation cues and scopes as additional entities with the corresponding entity types ("cue" or "xscope").
We have carried out experiments only with the training dataset (NegDDI-DrugBank Train 2013), i.e., training and testing on the 572 documents dataset (90% and 10%, respectively), and using the complete corpus (NegDDI-DrugBank 2013), i.e., the training dataset of the 2013 edition for training (i.e., 572 documents) and testing on the test dataset of the 2013 edition (i.e., 158 documents). For each of these experiments, we considered three situations: the original corpus without any negation annotations, the addition of only the negation cues and also considering the scope annotations. Results are presented in Table 8. Results show that we get different responses for each of the experiments when considering the negation annotations, nonetheless, both of them positive. When using only the training dataset, the number of true positives does not change much, but TEES returns less false positives, i.e, higher precision, without the degradation of the recall, thus, also with an improvement on the F-score. However, contrary to what was expected, the negation annotations had more effect on the recall for the 2013 test dataset, i.e., decrease of false negatives, instead on the precision, thus the additional true positives. A future error analysis on the results returned by TEES will shed some light on this behavior and give use some insights on how to better use negation annotations for drug-drug interactions.

7. Conclusions and Future Work

We have annotated the DDI-DrugBank 2013 corpus with annotations for negation following BioScope guidelines. It consists of 730 files with 6,648 sentences extracted from the DrugBank database. The extended corpus (NegDDI-DrugBank 2013) contains 1,448 sentences with at least one negation scope. This is the 21.78% of the sentences, confirming the tendency of use of negation expressions on biomedical documents.
We have computed correlations between the DDI and negation annotations present in the corpus. We can conclude that all these effective factors should be considered as potential features for a machine learning based method or in combination with a rule based system for extracting positive DDI from sentences with negation.
We plan to continue exploring the effect of features extracted from negation annotations in the DDI task, given the promising results which have been obtained in the preliminary experiments carried out with TEES, which explored only indirectly the potentials of negation cue and scope annotations.
Additionally, we plan to extend the annotations to the DDI-MedLine 2013 corpus. We expect differences in these annotations due of the language used in scientific publications.
Finally, we have used BioScope guidelines to annotate the NegDDI-DrugBank 2013 corpus but we plan to explore other guidelines as well, such as the one considered in *SEM conference (Morante and Blanco, 2012).

[7] http://bionlp.dbcls.jp/redmine/projects/bionlp-st-ge-2013/wiki/Wiki

8. Acknowledgements

The authors would like to acknowledge Prof. Ulf Leser (Humboldt-Universität zu Berlin) for the use of software resources. MN would like to acknowledge funding from the HPI Research School.

9. References

A.R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

M. Ballesteros, V. Francisco, A. Díaz, Herrera J., and P. Gervás. 2012. Inferring the scope of negation in biomedical documents. In *13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012)*, New Delhi. Springer, Springer.

J. Bjorne and T. Salakoski. 2013. Tees 2.1: Automated annotation scheme learning in the bionlp 2013 shared task. In *Proceedings of the BioNLP Shared Task 2013 Worksho*, pages 16–25.

B. Bokharaeian, A. Diaz, and M. Ballesteros. 2013. Extracting drug-drug interaction from text using negation features. *Procesamiento del Lenguaje Natural*, 51:49–56.

R. Bunescu, R. Ge, R.J. Kate, E.M. Marcotte, R.J. Mooney, A.K. Ramani, and Y.W. Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. 2002. Evaluation of negation phrases in narrative clinical reports.

Md. Chowdhury, M. Faisal, and A. Lavelli. 2013. Exploiting the scope of negations and heterogeneous features for relation extraction: A case study for drug-drug interaction extraction. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 765–771, Atlanta, Georgia, June. Association for Computational Linguistics.

J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. 2002. Mining MEDLINE: Abstracts, sentences, or phrases? In *Pacific Symposium on Biocomputing*, page 326. World Scientific Publishing Company.

J.H. Gurwitz, T.S. Field, J. Avorn, D. McCormick, S. Jain, M. Eckler, M. Benser, A.C. Edmondson, and D.W. Bates. 2000. Incidence and preventability of adverse drug events in nursing homes. *The American Journal of Medicine*, 109:87–94.

	true positive	false positive	false negative	precision	recall	F-score
NegDDI-DrugBank Train 2013	225	63	172	78.12	56.67	65.69
NegDDI-DrugBank Tain 2013+Cue	226	54	171	80.71	56.92	66.76
NegDDI-DrugBank Train 2013+Cue+scope	226	53	171	81.00	56.92	66.86
NegDDI-DrugBank 2013	602	173	278	77.67	68.40	72.74
NegDDI-DrugBank 2013+Cue	612	172	271	78.06	69.30	73.42
NegDDI-DrugBank 2013+Cue+Scope	618	175	265	77.93	69.98	73.74

Table 8: Results obtained from TEES for drug-drug interaction predictions using different configurations of the NegDDI-Drugbank corpus.

N. Konstantinova, S. de Sousa, N.P. Cruz, M.J. Maa, M. Taboada, and R. Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

J. Lazarou, B.H. Pomeranz, and P.N. Corey. 1998. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. *JAMA*, 279(15):1200–1205.

L.E. Martin. 1990. Knowledge extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*.

R. McDonald, F. Pereira, S. Kulick, S. Winters, Y. Jin, and P. White. 2005. Simple algorithms for complex relation extraction with applications to biomedical ie. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 491–498, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2012. *sem 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

R. Morante and W. Daelemans. 2012. Conandoyle-neg: Annotation of negation in conan doyle stories. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

A. Moreno Sandoval, S. Lopez Ruesga, F. Sanchez, and R. Grishman. 2003. Developing a spanish treebank. *Building and using parsed corpora*, pages 149–163.

C. Nedellec. 2005. Learning language in logic-genic interaction extraction challenge. In *Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*, volume 18, pages 97–99. Citeseer.

S. Pyysalo, A. Airola, J. Heimonen, J. Bjorne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9(Suppl 3):S6.

I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez. 2011a. Using a shallow linguistic kernel for drug-drug interaction extraction. *Journal of Biomedical Informatics*, In Press, Corrected Proof.

I. Segura-Bedmar, P. Martínez, and D. Sánchez-Cisneros. 2011b. The 1st ddiextraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. *CEUR-WS*, 761:1–9.

I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo. 2013. Semeval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, USA. Association for Computational Linguistics.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. Brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, April. Association for Computational Linguistics.

G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. 2008. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 38–45. Association for Computational Linguistics.

P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser. 2011. Relation extraction for drug-drug interactions using ensemble learning. In *Proceedings of the First Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*, pages 11–18.

D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. 2007. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*.

Semantic Relation Discovery by Using Co-occurrence Information

Stefan Schulz[1], Catalina Martínez Costa[1], Markus Kreuzthaler[1], Jose Antonio Miñarro-Giménez[2], Ulrich Andersen[3], Anders Boeck Jensen[4], Bente Maegaard[5]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria
[2]Medical Expert and Knowledge-Based Systems, Medical University of Vienna, Austria
[3]Sorano, Copenhagen, Denmark
[4]Center for Biological Sequence Analysis, DTU, Copenhagen, Denmark
[5]Institute for Language Technology, University of Copenhagen, Denmark

E-mail: stefan.schulz@medunigraz.at, catalina.martinez@medunigraz.at, markus.kreuzthaler@medunigraz.at, jose.minarrogimenez@meduniwien.ac.at, uan@sorano.dk, anders.boeck.jensen@gmail.com, bmaegaard@hum.ku.dk

Abstract

Motivated by the need of constructing a knowledge base for a patient-centred question-answering system, the potential of exploiting co-occurrence data to infer non-ontological semantic relations out of these statistical associations is explored. The UMLS concept co-occurrence table MRCOC is used as a data source. This data provides, for each co-occurrence record, a profile of MeSH subheading profiles. This is used as an additional source of semantic information from which we generate hypotheses for more specific semantic relations. An initial experiment was performed, limited to the study of disease-substance associations. For validation 20 diseases were selected and annotated by experts regarding treatment and prevention. The results showed good precision values (82 for prevention, 72 for treatment), but unsatisfactory values for recall (67 for prevention, 45 for treatment) for this particular use case.

Keywords: literature databases, knowledge engineering

1. Introduction

Motivated by a medical question answering use case we will investigate the automated construction of a supporting domain fact repository. Such a knowledge resource can be used for a series of possible applications, part of them directly related to biomedical text processing. In this context, the Danish ESICT project (Andersen et al., 2012) aims at developing strategies to provide natural language answers to laypersons' question on chronic diseases. One strategy of its hybrid approach has been the exploration of the content of SNOMED CT (2014), an ontology-based medical terminology. Its representational units (named SNOMED CT concepts, totalling about 300,000) group domain terms (about 700.000), which share a common meaning under specific concept categories (e.g. clinical finding, procedure, etc.) and are related by logical axioms, which express necessary and sufficient definitional conditions. Such axioms are strictly ontological, i.e. they state what is universally true for all individuals that instantiate a given SNOMED CT concept. Therefore, propositions of medical interest such as, for instance, relating typical diagnostic and therapeutic procedures, signs and symptoms to diseases, are not explicitly represented in SNOMED CT, which requires the exploitation of additional knowledge resources.

In this paper we investigate the potential of co-occurrence data as provided by the Unified Medical Language System (UMLS, 2014). They aggregate data on co-occurrences of semantic descriptors in bibliographic records. Their source is the MEDLINE database, in which each entry is indexed by a set of descriptors from the Medical Subject Headings (MeSH, 2014) vocabulary.

The goal of this study is to infer specific semantic relations out of these statistical associations.

Table 1 frames our hypothesis, viz. that co-occurrence patterns characterized by the defined semantic types involved, together with indirect semantic information of the context of a descriptor in the source (MeSH subheadings) may correspond to certain semantic relations. We focus on non-ontological predications, i.e. binary relations that are not used in definitions or universal statements as found in ontologies, thus excluding ontological relations like *part-of*, *has-site*, *has-quality* etc. In contrast, non-ontological predicates are not used in formal definitions as they express context-dependent and less strict associations, which, however, are often more "interesting" (Rector, 2008) from a medical point of view. Instead of stating what is necessarily true – like in formal ontologies – these predicates express what is typical, likely, or relevant. However, such knowledge is subject to change: A drug was indicated to treat a certain disease in the past, but it is mainly used for another purpose today or it has been withdrawn from the market. A clinical sign was frequently seen in the past, but it has become rare now, because the natural course of the underlying disease can no longer be observed, due to effective treatment.

Ideally, structured data in medical records would be a rich source of such associations. However, they are not

commonly available and often lag behind the state of the art of scientific investigation. Medical literature abstracts are much easier to access and their semantic metadata – in this case MeSH annotations – constitute a valuable source of knowledge. If we want to exploit them for knowledge construction, the question arises whether it is reliable to interpret semantic associations between topics in scientific literature in the light of clinical practice so that they can be used as a raw material for the acquisition of medical predications. This topic will be further discussed. The goal of this study is a first exploration of the possibility of constructing symbolic knowledge out of statistical associations. Whereas one could identify a much more complex matrix than Table 1, in this preliminary study we restrict ourselves to the exploration of disease-substance associations from where we attempt to extract knowledge on (i) how a disease can be treated, and (ii) how a disease can be prevented.

	Disease	**Finding**	**Substance**	**Organism**
Finding	sign of symptom of	accompanied by	treated by	affects caused by
Substance	*causes* *treats* *prevents* *metabolite*	causes treats prevents	interacts	affects is produced by
Organism	Causes affected by	Causes observed in organism	sensitive to	interacts with
Body part	possible location of	possible location of	targeted by	targeted by

Table 1. Examples of semantic relations between concepts ordered by semantic types. In this paper only co-occurrences between disease and substance concepts are explored (italics).

2. Materials

The Unified Medical Language System (UMLS) links representational units (classes, concepts, terms) from about 60 families of biomedical vocabularies. (Quasi-)synonymous terms are aggregated as UMLS concepts and identified by a unique identifier (CUI). The U.S. National Library of Medicine (NLM) produces and distributes the UMLS Knowledge Sources (databases) among which three files have been used in this work: (1) MRCOC, (2) MRCONSO and (3) MRREL.

- MRCOC provides pairs of concepts that co-occur in the same entries in some information source. It summarizes the MeSH descriptors that occur together in MEDLINE citations from the MEDLINE/PubMed baseline, a snapshot created at the beginning of each new MeSH indexing year. The co-occurrences are summarized by timeframe (MED, last five years of MEDLINE; MBD, previous five years of MEDLINE (years 6-10). In this study we have only used the most recent data set (MED). Besides the main headings, MRCOC includes a "fingerprint" of MeSH subheadings, such as "DT" (Drug Therapy) or "PC" (Prevention & Control). They characterize the context in which the first concept occurs in the related MEDLINE records. E.g., if an article is about the prevention of Stroke, the concept "Stroke" in the MEDLINE record is refined by the subheading "PC".
- MRCONSO relates UMLS CUIs with their language, source vocabularies, synonyms, translations, and lexical variants. We use this table to extract mappings between UMLS CUIs and SNOMED CT concepts. We are interested in SNOMED CT concepts due to their clinical relevance, but also because they are consistently grouped into semantic types, a few of them being depicted in Table 1.
- MRREL provides relation triples, again indexed by source. We will use this file for extracting hierarchical relationships.

Table 2 shows an example of the content of each of the previous files.

UMLS file	**Example of UMLS file content**
MRCOC	**C0000039**|**C0000506**|**MED**|L|**1**|**CH=1**|
MRCONSO	**C0000039**|**ENG**|S|L3000054|PF|S3260062|Y|A8383517|544223010|**102735002**|| **SNOMEDCT**|OF|102735002|**Dipalmitoyl-phosphatidylcholine (substance)**|9|O||
MRREL	**C0002871**|**CHD**|**C0002891**||MSH|MSH||

Table 2. The UMLS files used, with examples. The fields relevant for this work are picked out in bold face, such as concepts, co-occurrence frequency and subheading from MRCOC, the linkage to SNOMED CT in MRCONSO and a hierarchical relationship between two concepts in MRREL.

3. Methods

Our methods can be divided into:

- Mappings of co-occurrence pairs to SNOMED CT
- Calculation of the relative co-occurrence value
- Analysis of MeSH sub-headings
- Prototypical Implementation
- Evaluation strategy

Mapping of co-occurrence pairs to SNOMED CT

In order to add the corresponding SNOMED CT concept ID to each of the new UMLS concepts added, the UMLS MRCONSO file was used, which contains the SNOMED CT correspondences of the UMLS concepts when there is any. As there is a 1:n relation between UMLS and SNOMED CT concepts, mappings from numerous SNOMED CT concepts to the same UMLS concept occur. In MRCOC records we also find UMLS concepts that have no SNOMED CT correspondence. In such cases, the co-occurrence pair is discarded.

Computation of relative co-occurrence values

The overall frequency of MEDLINE annotations varies

across several orders of magnitude. A relatively low co-occurrence value may be more expressive than a higher one, in case the latter combines concepts of very high frequency such as, e.g. "Diabetes mellitus" and "Antibiotics". We hypothesize that this could be a source of error. We therefore used two thresholds, viz. the log likelihood ratio (Dunning, 1993) and an absolute threshold. A co-occurrence was seen as significantly expressive if the absolute co-occurrence was greater than five (McDonald, 2009), and the log likelihood ratio was greater than 6.63, which corresponds to $p < 0.01$.

Analysis of MeSH subheadings

After a thorough analysis of the 85 subheadings provided by MeSH we concluded that the subheadings "DT" and "PC" are highly indicative for the meaning of "treats" and "prevents", i.e. the two predicates we expect to induce. We decided to require the respective subheading for more than half of the co-occurrences. E.g., with a co-occurrence of 11, and the subheading distribution of "DT=5; PC=6", only "prevents" would be asserted.

Source concept	*Name (SNOMED CT)*	Bipolar disorder
	UMLS ID (CUI)	C0005586
	SNOMED ID	13746004
	SNOMED CT semantic type	Disorder
Target concept	*Name (SNOMED CT)*	Tricyclic antidepressant
	UMLS ID (CUI)	C0003290
	SNOMED ID	373253007
	SNOMED CT semantic type	Substance
MeSH subheadings		DT=9,CI=7,DI=5, PX=4,CO=2,EP=2, GE=2,BL=1,ET=1, PA=1,PC=1,PP=1, TH=1
Original Table	*Absolute co-occurrence*	17
	Relative co-occurrence (log-likelihood)	54.57

Table 3. Example record. The pairing of the two UMLS concepts has a co-occurrence in MED (last 5 year MEDLINE) of 17. If adding pairings from all UMLS subconcepts this value raises to 714. Relative co-occurrences are the decadic logarithms of the ratio of an absolute co-occurrence and a theoretic baseline (random co-occurrence). The MeSH subheading counts (for the original co-occurrence) refer to subheadings assigned to the first concept (or better the MeSH term mapped to it), e.g. nine DT (drug therapy), one PC (prevention and control)

Prototypical implementation

We developed a command line based interface in Java, using Apache Lucene 2.0 for indexing and the Apache Mahout Math library for log likelihood calculation. The *tempusfugit* package from Google Code was used for co-occurrence calculation. The interface provides a combined search of the fields shown in Table 3. The output result is shown in the console in descending order by relative co-occurrence in a short human readable format for quick search result feedback. Additionally, it is recorded as comma-separated values in an output file for further analysis. Further options include how many results should be returned as well as a cut-off-threshold for the co-occurrence value. VBA scripts were produced to facilitate the interpretation of the result by ordering, filtering, and colour coding the content in Excel spreadsheets.

Evaluation strategy

A reference standard was acquired by a random extraction of twenty disease terms from a textbook of internal medicine (Herold, 2014). For each disease, two of the authors, who are MDs, created a reference standard by collecting references to substances that are recommended for therapy and prevention. Various online and offline sources were used, with a focus on Wikipedia, as well as Danish treatment guidelines. Both MDs included only recommendations that appeared to be based on scientific evidence. For each disease, the co-occurrence table was filtered according to the following criteria, which had been heuristically acquired in a series of pre-tests with training data. The following parameters were measured, each for DT and PC:

Recall 1: Strict recall of reference standard concepts, regardless of hierarchical level.

Recall 2: Generous recall: here descendants of the concept in the reference standard were equally accepted.

Precision: Each retrieved concept is checked for correctness, i.e. whether it treats or prevents the disease under scrutiny. The criterion here is: given the state of the art, there is at least some clinical evidence that the treatment or prevention strategy is recommendable for humans. A thorough assessment would require a clinical review board. As this was not possible, we performed a cursory check of primary and secondary literature.

4. Results

Table 4 gives the result for the recall and precision analyses for all 20 sample diseases. Only for two diseases (*Infectious mononucleosis*, *Syncope*) the thresholds were not reached. More than ten results were only found for five diseases regarding therapy and two results regarding prevention.

In detail, the number of results per disease ranged for treatment from 0 to 40 (median = 2; mean = 6.7), and for prevention from 0 to 36 (median = 0; mean = 3.7). The strict recall values ranged from 0 to 1 (median = 0.35; mean = 0.42) for treatment and from 0 to 1 for prevention (median = 0. 50; mean = 0.49). The generous recall values

Disease	*# Target concepts*	*Recall (strict)*	*Recall (generous)*	*Precision (Correctness)*
Giant Cell Arteritis C0039483	13 / 0	1.00 / –	1.00 / –	0.77 / –
Cerebrovascular accident C0038454	40 / 36	0.50 / 0.57	0.83 / 0.86	0.62 / 0.83
Appendicitis C0003615	3 / 0	0.67 / –	1.00 / –	1.00 / –
Anthrax disease C0003175	1 / 2	0.10 / 0.30	0.10 / 1.00	1.00 / 1.00
Pre-eclampsia C0032914	6 / 6	0.50 / 0.33	0.50 / 0.33	0.50 / 0.16
Yellow fever C0043395	1 / 1	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
Gallbladder Carcinoma C0235782	3 / 0	0.33 / –	1.00 / –	1.00 / –
Membranous glomerulonephritis C0017665	10 / 0	0.67 / –	0.67 / –	0.90 / –
Hemolytic Anemia C0002878	2 / 0	0.33 / –	0.33 / –	1.00 / –
Hepatitis B C0019163	13 / 5	0.63 / 1.00	0.63 / 1.00	0.62 / 1.00
Impetigo C0021099	1 / 0	0.12 / –	0.12 / –	1.00 / –
Infectious mononucleosis C0021345	0 / 0	– / –	– / –	– / –
Pertussis C0043167	1 / 1	0.25 / 0.50	0.25 / 0.50	1.00 / 1.00
Malaria C0024530	14 / 16	0.36 / 0.67	0.36 / 0.67	0.79 / 0.75
Osteitis Deformans C0029401	2 / 0	0.22 / –	0.22 / –	1.00 / –
Neurosyphilis C0027927	2 / 0	0.20 / –	0.20 / –	1.00 / –
Gastric ulcer C0038358	19 / 7	0.22 / 0.00	0.22 / 0.00	0.53 / 0.00
Syncope C0039070	0 / 0	– / –	– / –	– / –
Tachycardia, Paroxysmal C0039236	2 / 0	0.50 / –	0.50 / –	1.00 / –
Erysipelas C0014733	1 / 0	0.25 / –	0.25 / –	1.00 / –

Table 4. Recall and Precision for DT ("Drug Therapy") / "Prevention & Control" (PC), with UMLS identifiers

(in which more general terms were allowed for matching) equally ranged from 0 to 1 (median = 0.35; mean = 0.45) for treatment and from 0 to 1 for prevention (median = 0. 77; mean = 0.67). Finally, the precision values ranged from 0 to 1 (median = 1; mean = 0.82) for treatment and from 0 to 1 for prevention (median = 0. 92; mean = 0.72). The large variation of the result is only understandable in a case to case analysis, which also reveals sources of error and demonstrates routes to improvement:

- Concept mismatch. Low recall values are often explained by the fact that the reference standard contained rather comprehensive lists of substance concepts (especially antibiotics), which were not contained in the co-occurrence table, to a large extent. A reference standard restricted to concepts that occur in the co-occurrence table would have yielded better results, as well as variations of the matching criteria, in the sense that a general term suggested by the system (e.g. *Antibacterial agent*) would match all existing antibiotics in the reference standard.
- Underspecifications of what was a therapy and what a preventive measure were observed in symptomatic treatments (e.g. *Intensive care treatment* in cases of *Yellow fewer*) or preventive measures that consist in the drug treatment of an underlying cause, e.g. treatment of *Arrhythmia* and *Arterial hypertension* as prevention of *Stroke*.
- Another issue is how to deal with widely practiced treatments of debatable evidence, such as, e.g. of

Glucosamine in *Osteoarthritis*.

- Interference with substance side effects. The zero precision at *Gastric ulcer* prevention was partly due to the fact that chemicals were found that *cause* gastric ulcerations. This phenomenon could only be observed here, because the sample did not contain other diseases that can be caused by substances. This shows the weakness of the subheading information we used, which was restricted to the analysis of the source concept only. The methods could be mitigated improved by a more detailed analysis of the subheading profile and the formulation of more rules, also including the relation "causes" as a possible outcome. In addition, the converse co-occurrence records could be included into the analysis.
- Lacking interest by the scientific community. The missing results for *Syncope* and *Impetigo* result from overall low (co-)occurrences. Treatment and prevention of these conditions have not changed for a long time, so that little information is contained in the co-occurrence dataset. In such cases, co-occurrence data from earlier time intervals could be useful. However, in case that co-occurrences of interest are only found in older datasets, a competing interpretation must be considered, *viz.* that a certain therapy became obsolete. For well-established, non-changing therapeutic measures, clinical co-occurrence data would probably yield less ambiguous results.
- No positive outcome of research. It is obvious that a high co-occurrence marked with an appropriate subheading profile will also occur in those cases where intensive research of drug effects could not be translated into clinical practice for several reasons, e.g. lack of superiority in clinical trials. An example found in our data was *Antiviral therapy* for *Yellow fever* or the use of *Zinc* in *Malaria* prevention. To identify such cases time series of co-occurrence data might be helpful, as well as filtering by publication types (e.g. *review* or *randomized controlled trial*).
- Ongoing research. This may produce high co-occurrence values even if a study is still restricted to animal models, such as current investigations of peptic ulcer prevention in rats using plant extracts. Here, information easily available in MEDLINE (*human* vs. *non-human*), but not connected to the co-occurrence dataset, could be used.

5. Related Work

Several authors have used the UMLS co-occurrence data, but there is a general impression that this resource has been rather underused. Burgun and Bodenreider (2001) analysed MRCOC, using three levels of semantic granularity, categorising by concept clusters that semantically cover a restricted area of interest. They found co-occurrence information helpful for this clustering task insofar as the redundancy between co-occurrence linkages and symbolic linkages were low. This corresponds to our division between ontological relations and non-ontological predicates. UMLS SN relationships could be relatively well inferred from co-occurrence information, like in our study, with a focus on the relation between disorders and chemicals. Question answering and enhanced information retrieval was a major driver of a study performed by Mendonça & Cimino (2000). They built their own co-occurrence table based on PubMed clinical queries and exploited its usefulness for gathering additional medical knowledge for knowledge base building. An automated approach for harvesting disease-chemical relationships was proposed by Zeng & Cimino (1998), based on UMLS MRCOC. They further evaluated the quality of the extracted knowledge by comparing the acquired relations with the expert system DXplain and manually extracted medical knowledge from literature. In disease-drug chemical relationships they achieved 93% sensitivity and 68% in disease-lab chemical relationships. Cantor et al. (2005) inferred gene-to-disease relationships using statistical and semantic relationships exploiting MRREL and MRCOC. Like in our approach they considered a threshold of five co-occurrence instances necessary to infer a concept-concept relation. They interlinked the relevant concepts with the Gene Ontology (GO) and evaluated the retrieved gene-to-disease relationships with the Online Mendelian Inheritance in Man's morbidmap (OMIM).

There are several systems that have successfully implemented information extraction methods to process biomedical literature databases. Examples are GOPubMed (Doms & Schroeder, 2005), MedlineR (Lin et all, 2004), FACTA (Tsuruoka et al. 2008), Alibaba (Plake et al., 2006), PolySearch (Cheng et al, 2008) and SemMedDB (Kilicoglu et al., 2012). All these systems concentrate on the knowledge extraction from title and abstracts using natural language processing methods. We have not found any previous work on the use of MeSH subheadings as an additional source of semantic information. The use of this information has been central in our work, which is, admittedly, preliminary and still restricted to the narrow scope of disease-substance associations. Another distinguishing feature of our approach is its reference to SNOMED CT as the emerging worldwide terminological standard. Although the conceptual space covered by MRCOC is much more coarse-grained, the availability of hierarchical links via the MRREL table allows inferring SNOMED-SNOMED co-occurrence values for concepts that have no direct representation in the MRCOC table.

The error analysis we performed on disease-substance co-occurrences demonstrated not only the need to refine the matching criteria but also to consider relations other than "treats" and "prevents", especially causation, we have ignored in this study, with the effect that the system suggested alcohol for prevention of peptic ulcers. Besides the need for improved criteria for the construction of the reference standard it has shed light on the (non-surprising) fact that research hypotheses and outcomes only partly translates in clinical practice.

6. Conclusion and future work

This study, motivated by the need to construct a knowledge base for patient-centred question-answering systems has been restricted to the investigation of substance-drug association, which is only one aspect. Next steps will be other semantic relationships as shown in Table 1, especially the relations between findings and diseases, with a focus on early diagnosis, risks, and prognostic factors. We have to keep in mind that the use of this kind of output can only be one of several knowledge sources in a question answering or decision support pipeline. Support by several sources and careful weighting are mandatory to prevent wrong answers or recommendations.

For the continuation of our work we still focus on MeSH annotations, and place special emphasis on the analysis of co-occurrences, but additional information from MEDLINE records such as timestamps, organisms, and publication types should additionally be exploited. This will require processing the whole body of MEDLINE. We will systematically analyse and categorise errors and try to identify indicative patterns for them. Currently we are incorporating co-occurrence based predications into a Web application, which compares several question-answering methodologies as developed by the ESICT project.

7. Acknowledgements

This paper was performed as a part of the ESICT project (Experience-oriented Sharing of health knowledge via Information and Communication Technology), funded by the Danish Strategic Research Council. http://cst.ku.dk/english/projekter/esict/.

8. References

Andersen, U., Braasch, A., Henriksen, L., Huszka, C., Johannsen, A., Kayser, L., Maegaard, B., Norgaard, O., Schulz, S., Wedekind, J. (2012). Creation and use of Language Resources in a Question-Answering eHealth System. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Burgun, A., Bodenreider, O. (2001). Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. *Studies In Health Technology And Informatics* 84, 171-175.

Cantor, M.N., Sarkar, I.N., Bodenreider, O., Lussier, Y.A. (2005). Genestrace - phenomic knowledge discovery via structured terminology. *Pacific Symposium On Biocomputing* 114, 103-114.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., Wishart, D.S. (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research* 36: W399-405

Doms, A., Schroeder, M. (2005). GoPubMed: exploring PubMed with the gene ontology. Nucleic Acids Research 33: W783-786.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Herold G. (2014) *Herold's Internal Medicine*, eBook at http://www.herold-internal-medicine.com/

Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., Rindflesch, T.C. (2012). SemMedDB: A PubMed-Scale Repository of Biomedical Semantic Predications. *Bioinformatics* 28(23): 3158-60.

Lin, S.M., McConnell, P., Johnson, K.F., Shoemaker, J. (2004). MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics* 20(18): 3659-61.

McDonald, John H. (2009) Handbook of biological statistics. Sparky House Publishing Baltimore, MD (2).

Mendonça, E.A., Cimino, J.J. (2000). Automated knowledge extraction from MEDLINE citations. *Proceedings of the Annual Symposium on Computer Application in Medical Care* 575-579.

Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., Leser, U. (2006) AliBaba: PubMed as a graph. *Bioinformatics* 22(19): 2444-2445.

Rector A. (2008) Barriers, approaches and research priorities for integrating biomedical ontologies. Available from: www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf.

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), 2008. Available from: http://www.ihtsdo.org/snomed-ct.

Tsuruoka, Y., Tsujii, J., Ananiadou, S. (2008). FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics* 24(21): 2559-2560.

Unified Medical Language System. (2014) http://www.nlm.nih.gov/research/umls/

Zeng, Q., Cimino, J.J. (1998). Automated knowledge extraction from the UMLS. *Proceedings of the AMIA Symposium* 568-572.

Building Realistic Potential Patient Queries for Medical Information Retrieval Evaluation

Lorraine Goeuriot[1], Wendy Chapman[2], Gareth J.F. Jones[1], Liadh Kelly[1], Johannes Leveling[1], Sanna Salanterä[3]

[1] Dublin City University, Ireland, [2] University of Utah, USA, [3] University of Turku, Finland

{lgoeuriot, gjones, lkelly, jleveling}@computing.dcu.ie, wendy.chapman@utah.edu, sansala@utu.fi

Abstract

To evaluate and improve medical information retrieval, benchmarking data sets need to be created. Few benchmarks have been focusing on patients' information needs. There is a need for additional benchmarks to enable research into effective retrieval methods. In this paper we describe the manual creation of patient queries and investigate their automatic generation. This work is conducted in the framework of a medical evaluation campaign, which aims to evaluate and improve technologies to help patients and laypeople access eHealth data. To this end, the campaign is composed of different tasks, including a medical information retrieval (IR) task. Within this IR task, a web crawl of medically related documents, as well as patient queries are provided to participants. The queries are built to represent the potential information needs patients may have while reading their medical report. We start by describing typical types of patients' information needs. We then describe how these queries have been manually generated from medical reports for the first two years of the eHealth campaign. We then explore techniques that would enable us to automate the query generation process. This process is particularly challenging, as it requires an understanding of the patients' information needs, and of the electronic health records. We describe various approaches to automatically generate potential patient queries from medical reports and describe our future development and evaluation phase.

Keywords: Query analysis, query generation, medical information retrieval

1. Introduction

Almost all over the world patient care in hospitals is documented carefully including admission information, documentation during the care episode and a discharge summary at the end of care. More and more often these documents are also shown or given to patients. They frequently include information that is not easy to understand by the patient, as they are written or dictated by a physician, nurse, therapist, specialist, or other clinician responsible for patient care. They describe the purpose of a hospital visit, completed procedures and investigations, the chosen treatment and care, a description of the recovery process, the status at discharge (discharge summary), and the future care plan. The primary purpose of discharge summaries is to support the care continuum as a handover note between clinicians, but they also serve legal, financial, and administrative purposes. The patient, her relatives, and other representatives are likely to have difficulties in understanding discharge text as in this simple example sentence from a US discharge: "AP: 72 yo f w/ ESRD on HD, CAD, HTN, asthma p/w significant hyperkalemia & associated arrythmias". These reports are written in specialised language, including abbreviations that are sometimes specific to individuals, as well as specialised vocabulary, that make them very hard to understand for patients (Allvin et al., 2011).

The overall, general objective of our research is to improve information access to health documents. Our research is conducted within within the ShARe/CLEF ehealth evaluation campaign (Suominen et al., 2013). The usage scenario of the campaign is to ease patients and their next-of-kin in understanding eHealth information. eHealth documents are much easier to understand after expanding shorthand, correcting the misspellings, normalizing all health conditions to standardized terminology, and linking the medical concepts to a patient-centric search on the Internet. The evaluation lab contains three tasks, the first one on visualization of eHealth data, the second one on information extraction from clinical data and the last one on patient-centered information retrieval. We are focusing here on the information retrieval task (Goeuriot et al., 2013a). Our main goal is to support patients in understanding their discharge summaries by understanding their information needs and automatically generating queries answering them. Although medical benchmarks for retrieving information related to queries exist, most of them focus on medical professionals' information needs rather than patients' needs. As commercial search engines query logs are generally not shared publicly, queries generated from patients' discharge summaries could provide essential material for patient-oriented novel applications, including generation of information to help patients understand their condition.[1] The automation of the query generation process would firstly greatly help the creation of new evaluation dataset. The more evaluation data there is, the more experiments there will be, and hopefully, the better medical IR system performances will be. Secondly, better IR results would also benefit medical professional: if the professionals could automatically generate queries and obtain relevant documents matching them from a medical report, this would greatly assist their patients in finding out about their condition.

In this paper we present a first step towards automatically creating these queries. After a description of related work (Section 2.), we identify the patients' information needs and detail them in Section 3. In Section 4. we describe the dataset (discharge summaries) and the annotations they

[1] Note that collections of medical documents typically contain sensitive (i.e. patient-related) information, which makes distribution even for research purposes difficult.

contain. In Section 5. we present the queries that have already been manually generated for the past evaluation campaign and the current one, and analyse them in the light of the identified patients' information needs. In the following section (Section 6.), we investigate possible techniques for automatic query creation and describe an evaluation framework for future experiments. Our conclusion and future work are presented in Section 7.

2. Related Work

McMullan (2006) provides a literature review on empirical studies on the use of the Internet for health information search by patients. The main outcome of that study is that the majority of health-related searches by patients target specific medical conditions. The search is carried out by the patient:

- before the clinical encounter to seek information to manage their own healthcare independently and/or to decide whether they need professional help;
- after the clinical encounter for re-assurance and deeper understanding or because of dissatisfaction with the amount of detailed information provided by the health professional during the encounter.

White and Horvitz (2013) study linking between patients' online behaviour and healthcare utilization. Their study is based on query logs obtained from a Microsoft browser toolbar and a survey. It shows a correlation between the query behaviour and the healthcare utilization. These findings fit our task scenario and confirm the patient need for information before and after being discharged from hospital.

To-date studies on patient queries have largely focused on query reformulation or query expansion. Their research is based on the following observations (Zeng et al., 2006):

- the patient queries are short;
- the queries often do not accurately reflect their information needs and are not effective for search.

Plovnick and Zeng (2004) conducted a pilot study where they reformulated consumer health queries with professional terms (UMLS preferred terms). While this method improved the results for queries containing acronyms and layperson terms, it seems to be mainly improving results from PubMed, which might not be the most searched and useful resource for patients. Zeng et al. (2006) describe the system HIQuA (Health Information Query Assistant), aiming at assisting users querying a search system to get health information. This system recommends additional or alternative query terms, and combines three sources: (1) usage patterns of consumers; (2) controlled medical vocabularies; (3) concept co-occurrence in medical literature. Their system recommendations resulted in statistically higher rates of successful queries, but had no statistically significant impact on user satisfaction or ability to accomplish predefined retrieval tasks.

Crain et al. (2010) skipped the interaction with the users, and instead created an information retrieval system based on Language Models and adapted to dialects. They mainly distinguish two dialects (common and technical language), but observed that the literature most often contains mixtures of dialects such as: common with some slang (on general discussion forums); common with technical (on laypeople health portals); and technical (on professional medical portals). Their system, called diATM is an extension of the polylingual topic models, that learn a language model in each language, for each topic. It shows improvement over state-of-the-art methods in providing relevant documents. While there has not been, to our knowledge, many studies on patient queries, the existing studies show that trying to link laypeople query terms to medical professional vocabulary and documents might not be the optimal approach. There is a need for studies which focus on patients' information search, and therefore a need for evaluation datasets.

3. Patients' Information Needs

Patients have several information needs when attempting to understand their discharge summaries. First, they would need to be able to read it, which means that acronyms should be expanded and normalized. Following this, the question of what the summary means arises. Often, patients would like to get more information about their disease: what it is, if it is dangerous, if it can be cured and what kind of changes and possible consequences it brings to their everyday life. Moreover, patients would need information about the possible treatments and discharge medication: what are the effects, the side-effects, is there any alternative treatment, etc. They might also need to know about their rights as patients, as well as contacts for support groups (Heikkinen et al., 2007). Their information needs also vary depending on their disease: patients with long-term diseases will have considerable knowledge related to the disease after a few years, while patients newly diagnosed or with short term diseases will have a very limited knowledge (Nuutila and Salanterä, 2006). Based on previous knowledge about patients typical information needs, we identified a list of information that we will focus on in this study:

1. Main disease(s) diagnosed
    - General Information: symptoms, risks, factors
    - Complications
    - Lifestyle
2. Treatment
    - General information: list of possible treatments
    - Surgery: procedure description, possible complications
    - Medication: description, precaution, side effects

4. Dataset Description

In this section we describe the dataset provided by the ShARe project, on which our query creation is based.

4.1. Discharge Summaries

The discharge summaries for the study were drawn from de-identified clinical reports originating from the ShARe corpus[2], which has added layers of annotation over a subset of the clinical notes in version 2.5 of the MIMIC II database[3]. The corpus contains 200 documents consisting of discharge summaries, electrocardiogram reports, echocardiogram reports and radiology reports, as described in Table 1. They were authored in an intensive care setting.

Table 1: Distribution of document type in the MIMIC corpus.

Type	# docs	per cent (%)
Discharge summary	62	31
Electrocardiogram report	54	27
Echocardiogram report	42	21
Radiology report	42	21
Total	200	100

The dataset used in this paper is a subset of MIMIC II, consisting of the 62 discharge summaries. The discharge summaries are semi-structured reports, they contain fields such as "admission date", "discharge date", "service", "medication", "allergies", etc. However, these fields can be missing or empty. Moreover, the discharge summaries contain many acronyms (e.g. "ICU", "HCT", "EGD", etc.), and highly specialised vocabulary (e.g. "dysphagia", "dysarthria", "ankylosing spondylitis", etc.). They may also contain spelling errors and typing errors. These characteristics make the data challenging to process, both from the point of view of the patient and from the point of view of the computer.

4.2. Annotations on the discharge summaries

Annotation of disorder mentions was carried out as part of the ongoing ShARe project. For this task, the focus was on the annotation of disorder mentions only. There were two parts to the annotation: 1) identifying a span of text as a disorder mention and 2) mapping the span to a UMLS CUI. Each note was annotated by two professional coders trained for this task, followed by an open adjudication step. UMLS represents over 130 lexicons/thesauri with terms from a variety of languages. It integrates resources used world-wide in clinical care, public health, and epidemiology. It also provides a semantic network in which every concept is represented by its CUI and is semantically typed (Bodenreider and McCray, 2003). A disorder mention is defined as any span of text which can be mapped to a concept in the SNOMED-CT terminology and which belongs to the Disorder semantic group. A concept is in the Disorder semantic group if it belongs to one of the following UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; Signs and Symptoms. The annotations cover approximately 181,000 words.

[2] https://www.clinicalnlpannotation.org

[3] Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5, http://mimic.physionet.org

5. Manual Query Generation

In this section, we describe the queries and how they were created for the previous eHealth evaluation campaign (2013) and for the current one (2014). An analysis of the 2013 query set is also provided, in the light of the patient needs highlighted in Section 3..

5.1. Topics from 2013

Our campaign provides an evaluation benchmark targeting patients search for medical information. An IR evaluation benchmark is generally composed of a document collection, a set of topics (extended queries), and relevance assessments (identifying for each topic which documents are relevant). Details on the 2013 benchmark can be found in (Goeuriot et al., 2013b).

Our interest here lies in the topics, which are extended queries. As we did not have access to any patient query logs, the decision was made to create queries from the discharge summaries (described in Section 4.1.), and especially to focus on disorders identified in them (described in Section 4.2.. For privacy reasons, we did not work with patients to generate the queries, but rather with experts in the domain, nursing researchers. These health professionals were provided with discharge summaries, in which one disorder had been randomly selected. Based on the contextual information in the report and the disorder picked, the health professionals created queries and additional fields forming the topics (following TREC standards):

Title: the text of the query;

Description: a longer description of what the query means;

Narrative: the expected content of the relevant documents;

Profile: a summary (such as age, gender, condition of the patient).

The task participants used these topics to evaluate their system performance, and were free to use external resources (such as medical terminologies, corpora, etc.), as well as the matching discharge summary as contextual information. The results and description of this 2013 task can be found in (Goeuriot et al., 2013a).

Although the same process was used to build each topic in the task, we observed differences among topics. These differences may be due to the fact that the topics are generated, from a highlighted disorder in a discharge summary, by a human estimating what the information need might be. Therefore, some topics may be directly related to a disease, while others enquire about the relationship between two disorders, or symptoms, for example. We thus categorize queries based on complexity, where complexity is derived from the number of concepts in a query. We define a concept as a specific medical entity (e.g. "diabetes

Table 2: Distribution of topic categories.

Category	# topics
1-concept	26
2-concepts	24
3-concepts	5
Total	55

mellitus" is a concept, "disease" is not). Based on this definition, we annotated the topics with the number of concepts they contained, based on the title (and if necessary, on the description).
The topic distribution for the annotated topics is shown in Table 2. We observe quite a balanced distribution between 1- and 2-concept topics.

5.2. Analysis of topics from 2013 in the light of the patients' information needs

In this section we describe a deep analysis of the queries and their content that has been conducted for this study, to gain a greater understanding of queries, towards automatic query generation. We annotated the queries according to the category of the concept they contained, i.e. disease, symptom, body part, treatment, procedure, care. Among the 1-concept ones, 96% belong to the disease category, which represents 25 queries, the remaining one relates to a body part. The 2-concepts categories distribution is shown in Table 3. Unsurprisingly, most of the multi-concept queries are centered on the link between a disease, and a concept from another category, the majority being another disease or a symptom. 3-concepts topics are more varied, and less numerous, so each of them contains a different combination of categories.
Single concept (1-concept) queries also quite often contain more general terms, that are not considered a concept, but more a facet of the main concept, for example "treatment", or "symptoms". We observed the distribution of these facets on the queries about a disease (N=25), reported in Table 4. When no facet was observed, we considered the patient to be looking for general information about the disease. The majority of the topics (15) have a facet with the disease. The facet can be the care, treatment or the symptoms of the disease. We also observed 4 queries relating to a disease in an acronym form. We considered this case independently, assuming that the patient would, besides general information, seek the meaning of the acronym.
While these queries have not been generated by patients, they give an idea of what a patients concern would be while reading their discharge summary. Moreover, they are very closely related to the information needs listed in Section 3.. While complex queries and relations between different entities have not been listed, we can see that, depending on the context and the knowledge the patient already has on his condition, it can be another information need.

5.3. Topics from 2014

Analysis of 2013 participating teams results showed that using the discharge summaries for contextual information did not improve IR system performance. One of the reasons explaining this is the way disorders within discharge summaries (from which the health professionals generated queries) were selected: they could either be one of the main disorders within the discharge summary, therefore be related to the discharge scenario; or be part of the patient history and mentioned in the discharge summary only for documentation, therefore not be related to the discharge scenario. Following discussion at the evaluation campaign workshop between participants and organizers, several solutions for building queries for the 2014 campaign were considered:

- Keep the topic creation based on a single disorder, therefore investigate ways to improve the selection of the disorder;
- Broaden the elements the topic could be based on: medication, surgery, care, etc.
- Consider the whole discharge summary as the topic/query.

The third option would make the task completely different from what it currently is, as it would require much more query processing: if the query is a full discharge summary, it would have to be processed first. Moreover, it would require a lot of preparation work, especially conducting the relevance assessment. As we would like the task to still be centered on the challenging improvement of health information retrieval, this option has not been further investigated.

Table 3: Distribution of categories in 2- and 3-concepts queries.

Categories	# topics
disease *AND* disease	10
disease *AND* symptom	5
disease *AND* body part	4
disease *AND* treatment	2
disease *AND* procedure	2
symptoms *AND* symptoms	1
disease *AND* disease *AND* disease	1
disease *AND* disease *AND* symptom	1
disease *AND* disease *AND* treatment	1
disease *AND* symptoms *AND* symptoms	1
symptom *AND* symptom *AND* symptom	1

Table 4: Distribution of facets in 1-concept queries relating to a disease category.

Facets	# topics
general information	10
acronym	4
care	4
treatment	4
symptoms	2
heredity	1

Table 5: Number of disorders listed in the discharge diagnosis field.

# Disorders	# Discharge Summaries	%
1	14	27
2	9	17.5
3	9	17.5
4	7	13
5+	13	25

Considering the first and second query generation methods, for time-related reasons, the first query generation method was selected for generation of the 2014 topics. A patient hospitalization is generally the cause of one or more disorders. These are often specified in the discharge summaries, or can be identified by health professionals. An analysis of the corpus described in Section 4.1. showed that among the 62 discharge summaries, 52 contained a field called *Discharge Diagnosis* or *Final Diagnosis*, containing a list of disorders the patient had been treated for. The number of disorders per discharge diagnosis is shown in Table 5. The topic creation has been made based on these identified diagnosis: if it contains only one disorder, the query is centered on it; if it contains more than one, the health professionals pick the disorder(s) they think the patient would first ask about. To avoid ending up with a biased set of topics, the health professionals were not given any specific guidelines for the disorder selection and quantity. As the task provides 5 training topics and 50 test topics, and only 52 discharge summaries contain a discharge diagnosis, it has been agreed with the health professionals that they would pick the main disorder from the remaining 3 summaries (in which all the annotated disorders were highlighted).
The structure of the topics remains the same as for 2013[4].
The second topic generation option, involving broadening the elements of the topic still has to be investigated. One approach to achieve this would be to combine our knowledge in patients' information needs and the information contained in the semi-structured and annotated discharge summaries to automatically generate patient queries. We present in the following section our primary investigations on this.

6. Towards Automatic Creation of Patients' Queries

6.1. Existing approaches

Ganguly et al. (2011) investigated generating queries automatically in the context of evaluation campaigns such as TREC 2010 session track, where the focus is the whole user session rather than single-queries. They reformulate existing queries, achieving either *specification*, where the reformulated query expresses a more precise information need, or *generalization*, where the reformulated query expresses a more general information need. They use a statistical corpus-based approach to retrieve more specialized or generalized terms to reformulate the query. As our main target here is to generate queries that are representative of patients information needs (over getting variants of existing queries), specification and generalization would not necessarily be the best approach. However, a corpus or ontology-based approach could allow identification of variants (e.g. synonyms, related disorders, or matching treatments) that could be used to generate query variants from an existing set. A risk would be the loss of the connection between the query and the discharge summary, and generating too homogeneous sets of queries.
Another strategy to automatically build queries from long documents involves summarizing them. Using classical summarization techniques based on statistical selection of segments in the text, a shorter version of a text (i.e. a query or document such as the discharge summary) can be generated to improve retrieval system performance (Arora et al., 2013). Using such a system in our case would require careful tuning of the summarization system, as classical summarization methods require adaptation to perform well on medical texts. This has been shown with scientific research articles (Nguyen and Leveling, 2013) and is even more the case with more condensed and less structured data such as medical reports.
A different approach would be based on the existing numerous work conducted on information extraction from electronic health records. Based on the existing annotations provided by the ShARe project (described in Section 4.2.), and the common semi-structure provided by the discharge summaries, information required to identify the main topics of interest of patients and generate queries from them. Based on the patients' information needs detailed in Section 3., the main fields that need to be identified in the documents are:

- the main disorders, often listed in the "discharge diagnosis" field of the discharge summaries;
- the treatment, often described in the "Discharge medication" field.

6.2. Evaluation Plan

We are planning on comparing these three methods. We will base our comparative analysis on two criteria:

- the quality of the generated queries, based on their relevance to the discharge summary, readability, and their usability;
- the quality of the results retrieved using standard IR systems: their relevance to the discharge summary, and the user satisfaction.

The quality of the generated queries will be manually assessed. Based on a given discharge summary, health professionals will rate generated queries according to their readability and relevance to the discharge summary. Information retrieval experts will judge how usable the queries for an IR evaluation task are, as our primary goal is to generate IR evaluation datasets.
As the set of queries for each discharge summary will be different depending on the method used, their relevance will be judged against the discharge summary (e.g. is that

[4]No example topic can be provided as the task is still running and only registered participants have access to the dataset.

query relevant to a patient receiving this discharge summary?). Similar to classical IR relevance assessment, health professionals will have to assess for each document its relevance to the discharge summary or a part of it.

7. Conclusion and Future Work

In this paper we investigated the generation of patient queries for an IR evaluation task. We first identified the patients' information needs, and then described how we manually created queries within CLEF eHealth. We then considered various ways to automatically generate queries, and how these methods could be adapted to medical IR: reformulation, summarization, and information extraction. We proposed a comparative evaluation plan to investigate these three approaches.

Our next step is to implement this comparative evaluation of the automatic patient query generation approaches. Another aspect worth exploring is the structure and the content of these queries, and the way this affects information retrieval performance. The evaluation campaign provides a privileged context to perform such experiments, with runs from various teams.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 257528 (KHRESMOI) and the ESF project ELIAS.

8. References

Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Daudaravicius, V., Hassel, M., Kokkinakis, D., Lundgren-Laine, H., Nilsson, G., Nytrø, O., Salanterä, S., Skeppstedt, M., Suominen, H., and Velupillai, S. (2011). Characteristics of finnish and swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics*, 2(Suppl 3).

Arora, P., Foster, J., and Jones, G. J. F. (2013). DCU at FIRE 2013: Cross-language !ndian news story search. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE) 2013*.

Bodenreider, O. and McCray, A. (2003). Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36:414–432.

Crain, S. P., Yang, S.-H., Zha, H., , and Jiao, Y. (2010). Dialect topic modeling for improved consumer medical search. In *Proceedings of AMIA Annual Symposium*, pages 132–136.

Ganguly, D., Leveling, J., and Jones, G. (2011). Automatic generation of query sessions using text segmentation. In *Proceedings of the Information Retrieval Over Query Sessions Workshop at ECIR 2011*.

Goeuriot, L., Jones, G. J. F., Kelly, L., Leveling, J., Hanbury, A., Mller, H., Salanter, S., Suominen, H., and Zuccon, G. (2013a). Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *CLEF online working notes*.

Goeuriot, L., Kelly, L., Jones, G. J. F., Zuccon, G., Suominen, H., Hanbury, A., Mller, H., and Leveling, J. (2013b). Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In Song, R., Webber, W., Kando, N., and Kishida, K., editors, *Proceedings of the 5th International Workshop on Evaluating Information Access (EVIA), a Satellite Workshop of the NTCIR-10 Conference*, Tokyo/Fukuoka, Japan. National Institute of Informatics/Kijima Printing.

Heikkinen, K., Leino-Kilpi, H., Hiltunen, A., Johansson, K., Kaljonen, A., Rankinen, S., Virtanen, H., and Salanterä, S. (2007). Ambulatory orthopaedic surgery patients' knowledge expectations and perceptions of received knowledge. *Journal of Advanced Nursing*, 60(3):270–8.

McMullan, M. (2006). Patients using the internet to obtain health information: How this affects the patienthealth professional relationship. *Patient Education and Counseling*, 63:24–28.

Nguyen, D. T. and Leveling, J. (2013). Exploring domain-sensitive features for extractive summarization in the medical domain. In *Proceedings of the Conference on Application of Natural Language to Information Systems (NLDB 2013)*, pages 90–101.

Nuutila, L. and Salanterä, S. (2006). Children with a long-term illness: parents' experiences of care. *Journal of Pediatric Nursing*, 21(2):153–60.

Plovnick, R. M. and Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology: a pilot study. *Journal of Medical Internet Research*, 6(3).

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G. K., Elhadad, N., Pradhan, S., South, B. R., Mowery, D., Jones, G. J. F., Leveling, J., Kelly, L., Goeuriot, L., Martínez, D., and Zuccon, G. (2013). Overview of the share/clef ehealth evaluation lab 2013. In Forner, P., Müller, H., Paredes, R., Rosso, P., and Stein, B., editors, *CLEF*, volume 8138 of *Lecture Notes in Computer Science*, pages 212–231. Springer.

White, R. W. and Horvitz, E. (2013). From health search to health care: Explorations of intention and utilization via query logs and user surveys. *Journal of the American Medical Informatics Association (JAMIA)*.

Zeng, Q. T., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., , and Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13(1):80–90.