Proceedings
of the 5th International Symposium on Languages in Biology and Medicine

12-13 DECEMBER, 2013

TOKYO, JAPAN

ISBN: 978-4-9907802-0-3

LBM Steering Committee:

Jong C. Park (KAIST, South Korea)

Limsoon Wong (NUS, Singapore)

See-Kiong Ng (I2R and SUTD, Singapore)

LBM 2013 is sponsored by:

DBCLS - Database Center for Life Science

KAISTCS - Department of Computer Science, Korea Advanced Institute of Science and Technology

LBM 2013 is supported by the journals:

JBMS - Journal of Biomedical Semantics

JCSE - Journal of Computing Science and Engineering

JBCB - Journal of Bioinformatics and Computational Biology

LBM 2013 is hosted by:

DBCLS - Database Center for Life Science Technology

LBM 2013 Homepage:

http://lbm2013.biopathway.org/

Printed in Tokyo by Database Center for Life Science – December 2013

Preface and Welcome Message

Dear LBM symposium participant,

It is a great pleasure to welcome you to the fifth symposium on languages in biology and medicine (LBM 2013) and introduce these proceedings

LBM is a biennial interdisciplinary forum that brings together researchers in biology, chemistry, medicine, public health and informatics to discuss and exploit cutting edge language technologies. Language, in its many forms, is the universal means to represent, convey, and question knowledge. Although knowledge is still widely communicated through natural languages, biology and medicine also use a number of other means of communication: sequences, ontologies, chemical and mathematical formulae, modelling languages, graphs, images, etc. Associated technologies such as text mining and information extraction, systems modelling, information visualization, semantic technology, and big data analytics are key for advancing biomedical research and healthcare provision. The automation and integration of all these solutions will enhance the access to the knowledge stored and conveyed in various representation, extending the opportunity of new knowledge discovery in the area. As all the individual technologies are constantly being challenged by user demands and complexities in an interdisciplinary research environment, the LBM symposium series aims to offer a forum for synergistic interactions between them.

LBM 2013, held on the 12th-13th December 2013, at the University of Tokyo, Japan, is the follow-up event of LBM 2011 (NTU, Singapore), LBM 2009 (Jeju Island, South Korea), LBM 2007 (Matrix, Biopolis, Singapore) and LBM 2005 (KAIST, Daejeon, South Korea).

A syster event, the International Symposium on Semantic Mining in Biomedicine (SMBM), has been held in 2012 at the Institute of Computational Linguistics, University of Zürich, Switzerland, in 2010 at EBI, U.K., in 2008 at the University of Turku in Finland, in 2006 at the University of Jena in Germany, and in 2005 at EBI, UK.

Submissions were invited in the following categories: full research papers, short papers, posters and highlight presentations. The latter category is an innovation meant to allow presentation of already published works to a potentially different audience, and thus broaden their impact.

A total of twenty five submissions were received. After a careful review process of each paper by at least two PC members (in most cases three), only two submissions were retained as full papers (out of 11 submitted in this category), seven were accepted as short papers, and eight were accepted as posters. Additionally, three submissions were accepted for highlight presentations.

We wish to express our gratitude as organizers and chairs to all the authors for the time and energy they invested in their research and for their choice of LBM 2013 as the venue to present their work. We are indebted to all members of the programme committee for their detailed inspection of all submitted work and their valuable comments. Additional thanks go to keynote and invited speakers for accepting our invitation and delivering inspiring talks.

Finally, we gratefully acknowledge all the work done by the members of the local organization committee, who invested a huge amount of time and energy to ensure the smooth running of this event.

Welcome again, and enjoy the symposium!

Fabio Rinaldi and Jin-Dong Kim (Program Chairs)

Jong C. Park, Limsoon Wong, See-Kiong Ng (Steering Committee)

Organization

Honorary Chair

Toshihisa Takagi (University of Tokyo, Japan)

General Chairs

Jong C. Park (KAIST, South Korea)
Limsoon Wong (NUS, Singapore)

Programme Chairs

Fabio Rinaldi (University of Zürich, Switzerland)
Jin-Dong Kim (Database Center for Life Science, Japan)

Local Organizing Chairs

Yasunori Yamamoto (Database Center for Life Science, Japan)
Hyunju Lee (GIST, South Korea)

Local Organizing Committee

Shinobu Okamoto (Database Center for Life Science, Japan)
Keiko Sakuma (Database Center for Life Science, Japan)
Kazuo Takei (Database Center for Life Science, Japan)
Pontus Stenetorp (University of Tokyo, Japan)

Programme Committee

Sophia Ananiadou (University of Manchester and NaCTeM, UK)
Eiji Aramaki (University of Kyoto, Japan)
Christopher Baker (University of New Brunswick, Canada)
Judith Blake (Jackson Lab, USA)
Olivier Bodenreider (National Library of Medicine, USA)
Wendy Chapman (University of Pittsburgh, USA)
Kevin Bretonnel Cohen (University of Colorado, USA)
Nigel Collier (NII, Japan)
Dina Demner-Fushman (National Library of Medicine, USA)
Juliane Fluck (SCAI, Germany)
Udo Hahn (Jena University, Germany)
Jörg Hakenberg (Icahn School of Medicine at Mount Sinai, USA)
Lynette Hirschman (MITRE, USA)
Chun-Nan Hsu (Academia Sinica, Taiwan)
Jaewoo Kang (Korea University, South Korea)
Arek Kasprzyk (Ontario Institute for Cancer Research, Canada)
Martin Krallinger (CNIO, Spain)
Michael Krauthammer (Yale University School of Medicine, USA)
Hyunju Lee (GIST, South Korea)
Hongfang Liu (Mayo Clinic College of Medicine, USA)
Peter Murray-Rust (University of Cambridge, UK)
See-Kiong Ng (I2R and SUTD, Singapore)
Jinah Park (KAIST, South Korea)
Sampo Pyysalo (University of Turku, Finland)
Dietrich Rebholz-Schuhmann (University of Zürich, Switzerland)
Rune Saetre (Norwegian University of Science and Technology, Norway)
Stefan Schulz (Medical University Graz, Austria)
Gerold Schneider (University of Zürich, Switzerland)
Adrian Shepherd (Birkbeck University of London, UK)
Jian Su (I2R, Singapore)
Yoshimasa Tsuruoka (University of Tokyo, Japan)
Ozlem Uzuner (State University of New York, USA)
Alfonso Valencia (CNIO, Spain)
Karin Verspoor (NICTA, Australia)
Maria Wolters (University of Edinburgh, UK)
Gwan-Su Yi (KAIST, South Korea)
Pierre Zweigenbaum (LIMSI-CNRS, France)

Program of December 12th, Thursday

08:45 - 09:15	Registration
09:15 - 09:30	Opening Session
09:30 - 10:30	Morning Session I
	KEGG molecular networks for linking genomes to society *Keynote* *Minoru Kanehisa*
10:30 - 11:00	Coffee break
11:00 - 12:30	Morning Session II
11:00	Hypothesis Generation in Large-Scale Event Networks *Kai Hakala, Farrokh Mehryary, Suwisa Kaewphan and Filip Ginter*
11:30	Distributional Semantics Resources for Biomedical Text Processing *Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski and Sophia Ananiadou*
11:50	Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction *Juan Antonio Lossio Ventura, Clement Jonquet, Mathieu Roche and Maguelonne Teisseire*
12:10	Open Information Extraction from Biomedical Literature Using Predicate-Argument Structure Patterns *Nhung Nguyen, Makoto Miwa, Yoshimasa Tsuruoka and Satoshi Tojo*
12:30 - 14:00	Lunch
14:00 - 15:00	Afternoon session I
	It's all semantics for the user *Keynote* *Martin Kuiper*
15:00 - 15:30	Coffee break
15:30 - 16:50	Afternoon Session II
15:30	Anatomical entity mention recognition at literature scale *Highlight* *Sampo Pyysalo*
16:00	Multilingual Annotation of Named Entities and Terminology Resources Acquisition (MANTRA) *Invited* *Dietrich Rebholz Schuhmann*
16:30	Sharing Reference Texts for Interoperability of Literature Annotation *Jin-Dong Kim*
16:50 - 17:00	Lightning Introduction to Posters
	TogoStanza: Semantic Web framework for SPARQL-based data visualization in the biological context *Shinobu Okamoto, Shuichi Kawashima, Takatomo Fujisawa and Toshiaki Katayama* Clinical Relation Extraction with Semi-Supervised Learning *Hiroki Ohba and Yutaka Sasaki* A Unique Linear Representation of Carbohydrate Sequences for the Semantic Web *Issaku Yamada and Kiyoko F. Aoki-Kinoshita* An Automatic Extractor for Biomedical Terms in Spanish *Leonardo Campillos-Llanos, José María Guirao-Miras and Antonio Moreno-Sandoval* OntoCloud – interactive visualization of relations between biomedical ontologies *Simon Kocbek and Jin-Dong Kim* A New Approach of Extracting Biomedical Events Based on Double Classification *Xiaomei Wei, Kai Ren and Donghong Ji* On Mention-Level Gene Normalization *Joonyeob Kim, Seung-Cheol Baek, Hee-Jin Lee and Jong C. Park* MPO: Microbial Phenotype Ontology for Comparative Genome Analysis *Shuichi Kawashima, Toshiaki Katayama, Toshihisa Takagi and Shinobu Okamoto*
17:00 - 18:00	Poster Session
18:00 - 20:00	Reception

Program of December 13th, Friday

09:00 - 10:00	Morning Session I
	Strategies for structuring free text to enable drug discovery and development *Keynote* *Phoebe Roberts*
10:00 - 10:30	Coffee break
10:30 - 12:00	Morning Session II
10:30	Incorporating Topic Modeling Features For Clinic Concept Assertion Classification *Dingcheng Li, Ning Xia, Sunghwan Sohn, Christopher G. Chute, Hongfang Liu and Kevin Bretonnel Cohen*
11:00	Vocabulary Expansion by Semantic Extraction of Medical Terms *Maria Skeppstedt, Magnus Ahltorp and Aron Henriksson*
11:20	Impact of real data from electronic health records on the classification of diagnostic terms *Alicia Pérez, Koldo Gojenola, Maite Oronoz and Arantza Casillas*
11:40	Comparison between Social Media and Search Activity as Online Human Sensors for Detection of Influenza *Mizuki Morita, Sachiko Maskawa and Eiji Aramaki*
12:00 - 13:30	Lunch
13:30 - 14:30	Afternoon Session I
13:30	Anatomography: an open anatomical mapping service for web-based healthcare communication and data visualization *Invited* *Kousaku Okubo*
14:00	OntoGene results in BioCreative and some thoughts about the nature of shared tasks in biomedical text mining *Highlight* *Fabio Rinaldi*
14:30 - 15:00	Coffee break
15:00 - 16:00	Afternoon Session II
15:00	Disease Gene Search Engine with Evidence sentences (version cancer) *Highlight* *Hyunju Lee*
15:30	Exploring sublanguages in biology and medicine *Invited* *Kevin Bretonnel Cohen*
16:00 - 16:30	Closing Session (Best Paper Award, SMBM 2014 announcement)
16:30 - 17:00	Free time
17:00 -	Excursion and Banquet (Sensoji Temple / Yakatabune Sumidagawa river cruise dinner)

CONTENTS

Posters

KEGG molecular networks for linking genomes to society

Minoru Kanehisa
Kyoto University
kanehisa@kuicr.kyoto-u.ac.jp

Abstract

The KEGG database that we have been developing since 1995 contains, among others, accumulated knowledge on metabolism, other cellular processes, organismal systems, human diseases and drugs represented as networks of molecular interactions, reactions and relations. The KEGG molecular networks, including KEGG pathway maps (graphs), BRITE functional hierarchies (ontologies) and KEGG modules (logical expressions), are widely used as a reference knowledge base for integration and interpretation of genome sequences and other types of data. In recent years the KEGG molecular networks have been expanded in two ways. One is linking genomes to phenotypes. The KEGG modules are being improved to automate interpretation of phenotypes, such as metabolic capacity and pathogenicity, from genome and metagenome sequences. The other is linking genomes to society. The KEGG MEDICUS translational bioinformatics resource has been developed by integrating drug labels (package inserts) used in society. The entire set of drug labels in Japan has also been processed to extract drug interactions associated with contraindications and precautions, as well as pharmaceutical additive and pharmacogenomic biomarkers. KEGG MEDICUS is directly targeted to society for helping to understand the scientific basis of diseases and drugs of personal interest.

Its all semantics for the user

Martin Kuiper
Norwegian University of Science and Technology (NTNU)
martin.kuiper@ntnu.no

Abstract

The Semantic Systems Biology group at the Norwegian University of Science and Technology is active in the field of biological knowledge management. The group has developed the BioGateway platform: a tool and resource set designed to provide the intended end-user, ideally a biologist or biomedical researcher, with assistance in the interpretation of her experimental data. The BioGateway platform is fueled by a data semantification pipeline that covers a chain of ontology manipulation, knowledge integration, pre-computing and reasoning, and data visualization. At various points along this pipeline we are active in making improvements with regard to performance, to keep the process tractable. Recently we published the orthAgogue software, boosting the speed of orthology prediction 200 fold. Even more recently, we ventured into a new approach to increase the speed of reasoning in order to eliminate a significant bottleneck in the generation of relational closures.

End users should, however, not be concerned with the details of this process, for all they care the technology can remain under the hood. What they do care about is an intuitive use of semantic resources and it is there that current interfaces for the design of SPARQL queries or inspection of the results leave a lot to be desired. Although for a query interface we have no improvements in preparation, for the visualization of ontologies we have produced the OLS Vis software, an intuitive and flexible viewer that is able to display complex ontologies.

The group is actively collaborating with end users, which resulted in the construction of a resource for gene expression regulation analysis: the GeXKB resource. Some examples of the use of GeXKB for regulatory pathway extension will be provided. The construction of GeXKB prompted us in the direction of semantifying data from the source: the curation of Transcription Factor information from scientific literature, resulting in the TFcheckpoint database (www.tfcheckpoint.org) and a set of curation guidelines for other volunteer curators to join in this effort.

We are now engaged in efforts to organize the global community interested in the domain of transcription regulation research to develop similar rigorous guidelines and apply them to develop an integrated homogeneous knowledge resource that could be used for instance in the field of gene regulatory network building and analysis. In order to further enable this work we are now configuring and using our literature curation environment SciCura (scicura.org, beta version) to facilitate the linking of proteins and genes, and their experimentally validated function with Uniprot / Entrez identifiers, and GO / PSI-MI / Brenda Tissue Ontology terms, respectively. We collect detailed information on gene regulatory events in the form of human-readable statements which are fully supported by proper database identifiers and ontology terms (an approach that is in fact extensible to create digital abstracts). These data will become publicly available and also form a crucial component of the GeXKB knowledge base.

References

Ekseth OK, Kuiper M and Mironov V. Orthagogue. 2013. an agile tool for the rapid prediction of orthology relations. *Bioinformatics*, Oxford Press, [Epub ahead of print].

Tripathi S, Christie KR, Balakrishnan R, Huntley R, Hill DP, Thommesen L, Blake JA, Kuiper M and Lgreid A. 2013. Gene Ontology annotation of sequence-specific DNA binding transcription factors: setting the stage for a large-scale curation effort. *Database*, Oxford Press, doi: 10.1093/database/bat062.

Chawla K, Tripathi S, Thommesen L, Lgreid A, and Kuiper M. TFcheckpoint. 2013. a curated compendium of specific DNA-binding RNA polymerase II transcription factors. *Bioinformatics*, Oxford Press, 29(19):2519-20.

Antezana E, Mironov V and Kuiper M. 2013. The emergence of Semantic Systems Biology. *New biotechnology*, Elsevier, 30(3):286-90.

A Venkatesan, V Mironov, M Kuiper. 2012. Towards an integrated knowledge system for capturing gene expression events. *Proceedings of ICBO 2012*, Graz, Austria, July 21-25.

Vercruysse S, Venkatesan A and Kuiper M. OLSVis. 2012. an Animated, Interactive Visual Browser for Bio-ontologies. *BMC Bioinformatics*, BioMed Central, 13:116.

Vercruysse S and Kuiper M. Jointly. 2012. creating digital abstracts: dealing with synonymy and polysemy. *BMC research notes*, BioMed Central, 5(1):601.

Strategies for structuring free text to enable drug discovery and development

Phoebe Roberts
Pfizer
Phoebe.Roberts@pfizer.com

Abstract

Therapeutic research creates scenarios that differ from academic research due to the constraints of linking targets, drugs and diseases for the sole purpose of treating patients. Pharmaceutical industry scientists routinely triage new drug target candidates, therapeutic modalities, and new indications for existing drugs. To do so effectively requires tapping into all available prior knowledge, be it public, subscription based, or internal to a specific organization. Prior knowledge represents the spectrum of findings derived from journal articles, patents or internal reports. The original free text represents one end of the spectrum from which entities and events are automatically extracted or manually curated. These derived statements increase in value when they are normalized to unique identifiers and codified in a machine-readable language. Accurate translation of free text into normalized, machine-readable complex statements is expensive, and we employ numerous strategies to meet the needs of biologists, chemists and clinicians looking for therapeutic opportunities. Investments range from licensing curated content and funding curation to hosting text mining systems, thereby ensuring comprehensive coverage and rapid turnaround.

Multilingual Annotation of Named Entities and Terminology Resources Acquisition (MANTRA)

Dietrich Rebholz-Schuhmann
University of Zürich
rebholz@ebi.ac.uk

Abstract

Mantra will provide multilingual terminologies and semantically annotated multilingual documents, e.g., patent texts, to improve the accessibility of scientific information from multilingual documents. The MANTRA project capitalizes on parallel document corpora from which translational correspondences will be computed by the use of different alignment methods. Fortunately, the biomedical domain, the application scenario of MANTRA, offers a rich variety of such parallel corpora.

The project partners will exploit these multilingual document sets to harvest terms and concept representations in different languages in order to augment currently available terminological resources such as the Medical Subject Headings (MeSH). The project partners will collaboratively build two types of resources:

- automatically enhanced multilingual terminologies and
- semantically annotated multilingual documents.

The novelty of the latter resource derives from the fact that we solicit and orchestrate community efforts for building up these annotated resources, a procedure that has already been proven successful for the semantic enrichment of large-scale biomedical document corpora (CALBC project) which was executed by the project partners.

The novelty of the first comes from a new combination of existing technologies in the area of statistical machine translation, named entity tagging and terminological resources. Both types of resources will be made available to the public for translation purposes and for search in and text mining from multilingual documents.

Source

Homepage of MANTRA project: http://www.mantra-project.eu/

Anatomography: an open anatomical mapping service for web-based healthcare communication and data visualization

Kousaku Okubo
National Institute of Genetics (NIG)
kokubo@genes.nig.ac.jp

Abstract

We report the world's first web-based anatomical mapping service for the public; it allows experts and non-experts to create and exchange three dimensional (3D) custom maps that offers common coordinate system through any web media. It consists of three elements: 1) map building kit: anatomically segmented parts of a digital 3D digital manikin, 2) map editor: a web-application that navigates the user to create custom map URL, 3) anatomical map API: a programming interface exposed to the web, on the image rendering server loaded with the kit data. The custom map URL, alone or inserted in source file of a web page, is visualized on the users browser as custom map image via map API. In addition, map API provides unique utilities such as anatomical address coding, collaborative mapping, anatomical choropleth maps that make the healthcare information more "actionable".

Exploring sublanguages in biology and medicine

Kevin Bretonnel Cohen
University of Colorado
kevin.cohen@gmail.com

Abstract

Sublanguages, or language in specialized domains and genres, are important to the understanding of biomedical languages and to the development of biomedical natural language processing systems. They also add to our understanding of language in general, and hence are important to corpus linguistics. In this talk, I present data from recent work on recognizing and characterizing biomedical sublanguages. It will be seen that current technologies are adequate for recognizing sublanguages in the biomedical domain: in scientific journal articles, in clinical documents, and in patents, as well as in languages with very different morphosyntactic characteristics from English. A toolkit for recognizing sublanguages is presented, and future directions in the characterization of sublanguages are discussed.

Highlight talk: Anatomical entity mention recognition at literature scale

Sampo Pyysalo and Sophia Ananiadou
National Centre for Text Mining and School of Computer Science, University of Manchester, UK

Anatomical entities are centrally important to biomedical discourse, and the ability to recognize mentions of these entities in free text is consequently required for comprehensive analysis of domain texts. Although there has been substantial progress over the last 15 years in the automatic recognition of mentions of various types of entities, relations, and events in biomedical scientific text, only limited effort has focused specifically on mentions of anatomical entities.

In this (proposed) highlight talk, we present our recent work "*Anatomical entity mention recognition at literature scale*" (Pyysalo and Ananiadou, 2013). The central contributions of this work are the following:

AnatEM corpus The extended Anatomical Entity Mention (AnatEM) corpus is a manually annotated corpus of 1212 PubMed abstracts and PMC Open Access subset full-text extracts (approx. 250,000 words). The corpus annotation extends substantially on previously available resources (Ohta et al., 2012), identifying over 13,000 mentions of anatomical entities and assigning each to one of 12 granularity-based types such as CELLULAR COMPONENT, TISSUE, and ORGAN (Figure 1). The corpus is made available under the open Creative Commons BY-SA license.

AnatomyTagger We evaluated numerous strategies to improve the performance of machine learning-based anatomical entity mention recognition, including dictionary resources based on UMLS and OBO Foundry ontologies, statistical truecasing, and incorporation of non-local features through multi-stage tagging. The most effective strategies were then implemented in AnatomyTagger, a standalone tagger using the NERsuite toolkit,[1] which is in turn built on the CRFsuite (Okazaki, 2007) implementation of conditional random fields.

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

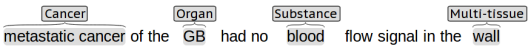


Figure 1: Annotation example

Evaluation on the AnatEM corpus showed that the AnatomyTagger outperforms both dictionary-based approaches as well as several recently proposed machine learning-based taggers, achieving an F-score of 92% for mention detection and 85% for detection and classification. The tagger is released under the open source MIT license.

Literature-scale application We applied an implementation of the AnatomyTagger pipeline in the UIMA system to automatically tag all 600,000 PMC Open Access full-text documents, resulting in the recognition of over 48 million anatomical entity mentions. The resulting dataset, representing the first application of a machine-learning based anatomical entity recognition system to the entire Open Access biomedical literature, opens several new opportunities for detailed analysis of the scientific literature. This dataset is made available under the open Creative Commons BY-SA license.

Availability All introduced tools and resources are available from `http://nactem.ac.uk/anatomytagger/`

References

T. Ohta, S. Pyysalo, J. Tsujii, and S. Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of DSSD 2012*.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs).

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*. (in press).

OntoGene results in BioCreative and some thoughts about the nature of shared tasks in biomedical text mining

Fabio Rinaldi
University of Zürich
fabio.rinaldi@uzh.ch

Abstract

Since 2003 the BioCreative challenge has addressed various aspects of text mining that could be relevant for the curation of biological databases, such as Gene Ontology annotations, detection of gene mentions, database normalization of gene mentions, protein-protein interactions, up to experimental interactive curation. In a parallel development, the BioNLP shared task has addressed the problem of finding biomolecular events which appear in the biomedical literature, provided some evidence for the entities involved in the event. Although these problems might appear as distinct, they are related by the fact that in a complex real-world curation pipeline they would need to be combined in order to provide a complete text mining solution in support of detailed curation needs.

Starting in BioCreative 2012, and again in BioCreative 2013, the Comparative Toxicogenomics Database (CTD) has organized a task specifically designed to support their daily curation process, providing its own curated data as training and test sets. I will present the method and results used by the OntoGene team [1] for their participation in the CTD task [2], and use it as an example in order to discuss the different nature of the two main competitive evaluations in the biomedical text mining domain (BioCreative and BioNLP), as well as the different way in which training/test data is obtained, which I think is a crucial point for the relevance of the results.

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

[2] http://database.oxfordjournals.org/content/2013/bas053.full

Short Bio

Fabio Rinaldi is the leader of the OntoGene research group at the University of Zurich and the principal investigator of the SASEBio project (Semantic Enrichment of the Biomedical Literature). He holds an MSc in Computer Science and a PhD in Computational Linguistics. He is author or co-author of 100+ scientific publications (including 19 journal papers), dealing with topics such as Ontologies, Text Mining, Text Classification, Document and Knowledge Management, Language Resources and Terminology. Currently his main research area is biomedical text mining.

DigSee: disease gene search engine with evidence sentences (version cancer)

Jeongkyun Kim[1], **Seongeun So**[1], **Hee-Jin Lee**[2], **Jong C. Park**[2], **Jung-jae Kim**[3] and **Hyunju Lee**[1,*]

[1] School of Information and Communications, Gwangju Institute of Science and Technology (GIST)
[2] Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST)
[3] School of Computer Engineering, Nanyang Technological University (NTU)
*hyunjulee@gist.ac.kr

Abstract

Biological events such as gene expression, regulation, phosphorylation, localization and protein catabolism play important roles in the development of diseases. Understanding the association between diseases and genes can be enhanced with the identification of involved biological events in this association. Our novel search engine, DigSee, services the sentences with those identified triple relations, on the requests from users, which require information such that which genes are involved in the development of which disease through which biological events.

In DigSee, the candidate evidence sentences are ranked based on a Bayesian classifier to measure the relevance of the sentences, which means whether the recognized gene is the subject of the identified event that leads to changes of the given diseases properties. The classifier uses 10 linguistically motivated features, including features obtained from dependency parse trees, and handcrafted cancer-related terms, and terms related to negative sentences. The model is trained and tested on a gold-standard data set manually constructed by the authors. In the current version of DigSee, 1 391 019 evidence sentences from cancer-related MEDLINE abstracts were collected, and DigSee supports all cancer types (200 cancer names) and the following event types as the molecular context of genedisease association: gene expression, transcription, phosphorylation, localization, regulation, binding and protein catabolism. DigSee is available through http://gcancer.org/digsee.

References

Jeongkyun Kim, Seongeun So, Hee-Jin Lee, Jong C. Park, Jung-jae Kim and Hyunju Lee 2013 DigSee: disease gene search engine with evidence sentences (version cancer) *Nucleic Acids Research*, Oxford Press, Vol. 41, Web Server issue, doi:10.1093/nar/gkt531, W510-W517.

Hypothesis Generation in Large-Scale Event Networks

Kai Hakala[1], **Farrokh Mehryary**[1], **Suwisa Kaewphan**[1,2], **and Filip Ginter**[1]
[1]University of Turku, Turku, Finland
[2]Turku Centre for Computer Science (TUCS), Turku, Finland
first.last@utu.fi

Abstract

Hypothesis generation from literature is among the most prominent goals of the BioNLP research community. The existence of EVEX, a large-scale event network mined from the entire available biomedical literature, opens the possibility to cast this task in a supervised machine learning setting, defining it as the prediction of edges in this network, based on features from their network context.

In this paper, we study the task from two perspectives. First, we build a machine learning system which predicts novel pairwise relationships in the EVEX network and evaluate its performance using both the standard measures as well as through a manual inspection on a subset of the output. And second, we analyze and discuss the issues in evaluation arising from cross-validation in densely connected graphs with uneven edge distribution.

We find that the task is learnable, achieving performance clearly above baseline. Further, a manual inspection of predictions not found in the EVEX network showed several candidate pairs, whose interaction could be verified in the literature. These pairs hint at the possibility that true novel interacting pairs were identified by the system as well, even though further work is necessary to confirm whether that is indeed the case.

1 Introduction

Hypothesis generation based on literature mining is among the most prominent goals of the BioNLP research community. Already over 20 years ago, the legendary ARROWSMITH system (Swanson, 1988) identified novel association candidates by combining the information from entity pairs frequently co-occuring in the literature (Bekhuis, 2006). The work of Swanson, and many others, was based on the statistics of term co-occurrence in text. To increase the recall of low-frequency associations, subsequent work has focused on a more detailed extraction of pairwise interactions of (mainly) genes and proteins from individual sentences (Pyysalo et al., 2008; Tikk et al., 2010). Such extraction of interacting pairs has the advantage that even single assertions can be extracted, without the need for sufficiently high co-occurrence. These methods are, however, often largely restricted to the extraction of untyped, undirected pairs, i.e. an association is postulated, but no additional knowledge regarding its type is given. Finally, methods for the extraction of detailed *events* have been introduced, mainly as the outcome of the BioNLP Shared Tasks on Event Extraction (Kim et al., 2009; Kim et al., 2011; Nédellec et al., 2013). The events are detailed, recursive structures that provide a more faithful representation of the semantics of the underlying text. Event extraction systems have subsequently been applied on a large scale to the collection of PubMed abstracts and the open-access section of PubMed Central full-text articles (Björne et al., 2010; Gerner et al., 2012). EVEX (Van Landeghem et al., 2013), presently the only publicly available large-scale event collection, serves as the basis of this study and is discussed in more detail in Section 2.

The availability of EVEX as a large-scale network, with genes and gene products (GGPs) as the nodes and their relationships as the edges, allows us to study the problem of hypothesis generation at a large scale and in a machine learning setting. Rather than relying on a set of pre-defined patterns, such as the triangular pattern used by Swanson which postulated the hypothetical association

A–C given the identified associations *A–X* and *X–C*, we define a number of features extracted from the network context and train a supervised classifier. This allows us to incorporate more information into the classification process.

Given a candidate pair of nodes not already connected by an edge in the network, the task is to predict the existence of a potential edge, or edges, between the two nodes and possibly also the nature (type) of the predicted relationship. Features for this prediction task are extracted from the existing network neighborhood of the candidate pair, in particular from short paths in the network that connect the two nodes. Edges already existing in the network are then used as positive examples in training. In this paper, we will explore both the simpler task of predicting whether an edge exists or not, as well as the more complex multi-label task of predicting also the type of the newly predicted edges.

2 Data

The data we use is extracted from EVEX, a large-scale literature mining resource built on top of the set of events extracted from all PubMed abstracts and PubMed Central Open Access full-text articles, using the TEES system (Van Landeghem et al., 2013; Björne et al., 2012). A feature of EVEX particularly important for this current study is that it provides a *network view*, where GGPs are normalized to their respective Entrez Gene identifiers using the GenNorm system (Wei et al., 2012), and the complex recursive events are reduced into pairwise relationships with the coarse-grained types of *Regulation, Binding*, and *Indirect regulation* and 29 fine-grained types such as *Regulation of phosphorylation* and *Indirect catalysis of hydroxylation*. This network view thus abstracts away some of the complexity of the recursive events and allows modeling the problem as a simple edge prediction in directed graphs. Figure 1 illustrates a tiny part of the human gene regulatory network extracted from the EVEX resource.

In the EVEX *network view*, the individual event occurrences extracted from text are aggregated, i.e. a single edge in the network stands for all individual events that represent this relationship anywhere in the literature. This is possible because the GGP symbols are normalized into Entrez Gene identifiers and all edges of the same type and direction between the same Entrez Gene identifier pair can be merged. This has the major advantage of allowing the use of features from all the available literature when predicting new relationships, not restricting ourselves to a single sentence, or a single article.

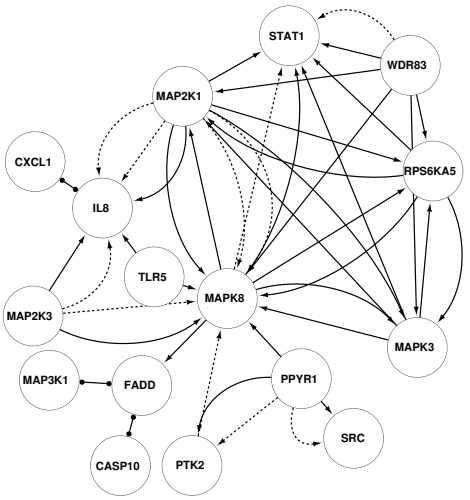


Figure 1: A tiny part of the highly connected network extracted from EVEX for human gene/protein interactions. Circle-terminated connections indicate binding and arrows indicate regulation. Indirect regulations are presented with dashed lines while direct regulations are presented with solid ones.

The complete EVEX network consists of 819,348 unique edges among 48,061 unique GGPs from a large number of different organisms. To deal with a smaller, yet biologically motivated problem for this initial study, we selected the subnetwork formed by all human genes (judged by their Entrez Gene identifier) and only consider the three coarse-grained types, rather than the 29 fine-grained types available in EVEX. This human gene network consists of 13,418 nodes and 265,738 directed edges. As illustrated in Table 1, the network is densely connected, with 97.6% of nodes belonging to a single large connected component. To simplify processing, we remove the 317 nodes that belong to connected components with less than 8 nodes, and the 76 edges among these nodes. The 212 connected components with only a single node reflect the self-interacting genes with no known interactions with other genes in the EVEX database.

# nodes	# components
1	212
2	38
3	6
4	2
6	1
7	1
13,091	1

Table 1: The distribution of connected components in the network, showing that essentially the entire network is spanned by a single connected component with 13,091 nodes.

3 Methods

Casting the task in a straightforward supervised machine learning setting, we need to specify what our positives and negatives are. A positive example is a pair of nodes in the network which is connected with an edge. In the classification, we will use features extracted from paths two or three edges long that connect the two nodes in the network.

As with many similar problems, there is no apriori given set of negative examples. Instead, any pair of nodes that is not directly connected in the network can be technically considered as a negative example. This would, however, have two unwanted consequences: First, the number of such negative examples would be enormous in comparison to the number of positive examples, and second, most arbitrary node pairs are distant in the network and obviously unrelated. The classification problem would thus become trivial if trained and evaluated on such negative examples, and its performance would not be very informative. Rather, we thus restrict the selection of negative examples to the "interesting" node pairs that are not connected by a direct edge in the graph, but are connected by at least one path of at most three edges. In this way, we focus on the more realistic problem of predicting novel relationships for node pairs that are closely connected in the network.

Current state-of-the-art event extraction systems perform in the range of 40–50% in terms of recall. Due to this fairly low retrieval rate some of the examples labeled as negatives in training are in fact false negatives in the underlying EVEX network, and are bound to add noise to the training and evaluation data. To diminish their effect, we further refine the data by excluding negative gene pairs that co-occur in at least one sentence. Since, as was shown for example in the Genia Shared Task data, statements of interactions rarely cross the sentence boundary, this filtering step will remove most of the EVEX false negatives. The final set of negatives used in training and evaluation is thus constituted by pairs that are connected in the network by at least one path of length at most three edges, and that have not co-occurred in a sentence.

Comparing the average number of paths in the network that connect candidates in the final set of positives (32,077), the final set of negatives (427), and the (currently discarded) set of negatives where the candidate GGPs co-occur in a sentence (8,602), reveals large differences, in particular further confirming that the currently discarded negatives probably contain a non-trivial proportion of actual existing interactions that the EVEX text mining system failed to extract. Even though these examples are excluded in the current evaluation so as to avoid the added noise in the data, future work should focus on assessing their importance in hypothesis generation as well as in improving the recall of the EVEX resource.

4 Features and Classification

To solve the binary classification problem of predicting the existence of an edge, we train a linear support vector machine using the SVM-light library (Joachims, 1999). The features used are based on the paths between the nodes, limiting to only the paths of length two and three. Two feature types are used:

1. For every unique path type, defined as the concatenation of edge types and directions along the path, the number of paths of this type connecting the pair of GGPs is given.

2. For every unique path type of length two edges, the maximum of EVEX confidence scores of the edges in the path. The confidence scores given in EVEX for the individual edges reflect the reliability of the underlying events being correctly extracted from the text.

The first set of features is purely based on the structure of the graph and could be used with various graphs constructed from different data sources. The second set, however, is unique to the underlying text mining resource, providing information that cannot be acquired from other type of

gene regulatory networks. The performance gain of these feature types is discussed in Section 5.3. It is worth noting that neither of these feature types encode information about the intermediary nodes in the paths nor the textual context where the interactions have been seen. As will be discussed in detail in Section 5.2, this is particularly important in the cross-validation setting, where it is difficult to avoid paths crossing between training and testing sets without substantially changing the characteristics of the data.

The optimization of the classifier C parameter was done with a grid search against a development set. As the natural distribution of positive and negative examples is very tilted, we oversample positive examples to create training data with a 1:10 proportion between positive and negative examples. No such oversampling is done for the development and test sets, naturally.

The more complex problem definition, where event types are also predicted, can be formalized as a multi-label classification task. In this case we use a one-vs-all classification approach implemented with the scikit-learn library (Pedregosa et al., 2011) and a linear support vector machine. The same features are used in both tasks.

5 Results

5.1 Baseline

Even though we select the negatives to be connected with a path of at most three edges, there is still a clear difference in the density — i.e. the number of paths in the network that connect the the two nodes — between the positive and the negative examples. The positive examples have on average a notably higher number of connecting paths. The distributions of the path counts are illustrated in detail in Figure 2. The histograms show that the distribution of the negative examples resembles an exponential distribution whereas the positive examples show a heavy-tailed distribution. This is naturally a difference which a classifier can learn to exploit. To test how predictive the path count is of the classification outcome, we train a baseline classifier which is only given the total number of connecting paths.

5.2 Test Protocol

All experiments are carried out using the 10-fold cross-validation protocol, whereby the network is split into ten sub-networks, of which eight are used for training, one for parameter optimization, and one for testing. 20,000 pairs with at least one connecting path of at most three edges are randomly sampled from each partition to form test sets with a natural distribution of positive and negative examples. The results on the ten sets are then averaged. Unlike in most machine learning problems where individual instances are largely independent, the densely connected event network complicates the 10-fold split substantially. The obvious approach of splitting the nodes randomly is not practical because for any given node, 90% of its neighbors will be assigned to a different set than the node itself, while for feature generation and testing it would be desirable for the node as well as its neighbors to be assigned to the same set. Rather than splitting the nodes into sets randomly, we apply the METIS toolchain for graph partitioning (Lasalle and Karypis, 2013), which heuristically splits the network into roughly equally-sized parts while minimizing the number of edges crossing among the parts.

An issue with splitting the graph into partitions roughly equally-sized with respect to the number of nodes is that a small number of extremely densely connected hub nodes causes large variations in edge density in the resulting sets and, as will be shown later, subsequent variation in the results. The METIS algorithm allows for weights to be given to the nodes, which affects the division to create graph partitions with roughly equal sum of node weights. Weighting the nodes by their degree, we can thus subdivide the graph into partitions with a roughly equal number of edges, thus balancing the edge density rather than node density. To illustrate the difference, in Table 2 we show the number of nodes and edges in two METIS-based 10-fold splits corresponding to the two aforementioned strategies. Note the particularly disturbing fold no. 1 in the unweighted strategy, which has an order of magnitude more edges than any of the other nine folds. This partition includes several well-studied genes such as TNF-alpha, IL-6 and insulin, all with hundreds of known interaction partners in the EVEX resource.

Another problem stems from the fact that, regardless of the strategy used to divide the nodes into subsets, there will be a number of edges spanning across these subsets. Of particular concern are edges spanning between the training and the test set in a given fold of the 10-fold protocol.

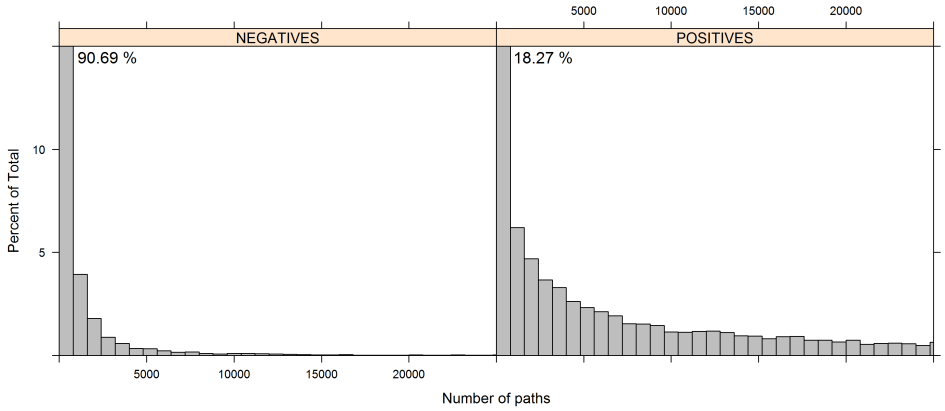


Figure 2: Distributions of positive and negative examples in terms of the connecting path counts. The y-axis has been limited to 15% and the actual heights of the bins exceeding this limit are denoted in the figure.

	Unweighted		Weighted	
Fold	Nodes	Edges	Nodes	Edges
0	1,335	3,713	904	7,664
1	1,348	79,005	1,287	9,678
2	1,320	4,263	1,892	7,677
3	1,348	9,959	1,369	8,326
4	1,302	3,198	1,256	11,616
5	1,278	1,738	1,140	8,668
6	1,289	2,129	921	8,533
7	1,296	2,273	1,792	8,333
8	1,283	3,116	1,180	6,786
9	1,292	2,221	1,350	8,421
Total	13,091	111,615	13,091	85,702

Table 2: The distribution in terms of the number of nodes and edges when splitting the network into 10 folds with roughly equal node count (unweighted) and roughly equal edge count (weighted) using the METIS algorithm.

While the obvious "safe" course of action would be to remove all edges that connect nodes between the training and test data, this has a notable impact on the data exactly because it is so densely interconnected. This is again illustrated in Table 2, which shows the edge counts for the 10 partitions when edges spanning across partitions are removed. In the unweighted set, 58% of edges are removed, and in the weighted set, full 68% of edges are removed. Removing edges spanning across the 10 folds thus clearly substantially affects the properties of the underlying data. Note that while only removing edges spanning between the training and test set in every iteration of the 10 fold evaluation strategy is also an option, this would result in substantially skewed distributions between the training and test data, and we thus do not consider this approach further.

To assess the impact of these choices, we carry out evaluation on all four combinations, i.e. splitting to balance the number of nodes versus number of edges, and preserving or removing the edges spanning between the folds in the 10-fold protocol. The four resulting divisions and their salient characteristics are summarized in Table 3.

5.3 Classification Results

In the evaluation, we compare classifiers with three different feature sets on the four network partitioning strategies introduced in Section 5.2. The baseline classifier uses only one feature which encodes the total number of paths connecting the candidate pair. A second classifier utilizes counts of unique path structures, and a third classifier introduces also features encoding confidences of the individual edges, as extracted from EVEX. Precision, recall, and F-score averaged over the 10 folds for the three classifiers are shown in Table 4.

Several observations can be made: To begin with, the performance of the baseline classifier is very poor in evaluation strategies with equal node count partitioning, achieving F-scores of 7.36% and 3.65%. This is most visible when edges spanning across partitions are retained in the data, where the baseline classifier obtains an F-score of 0.0 in four folds out of ten. This is likely because the baseline classifier can only rely on the number of paths, which substantially differs among

Dataset name	Positives Frequency (%)	Negatives Frequency (%)	Average paths count (Total)	Average paths count (Positives)	Average paths count (Negatives)	Sample STD (Total)	Sample STD (Positives)	Sample STD (Negatives)
unweighted/remove	3.86	96.14	1289.00	12590.00	446.20	10325.95	42184.00	3267.35
weighted/remove	1.72	98.28	195.40	3086.00	104.40	929.10	5134.79	331.40
unweighted/keep	3.86	96.14	1543.00	16480.00	943.10	11358.70	52216.45	3915.70
weighted/keep	1.72	98.28	1094.00	25050.00	673.60	7932.11	52306.50	2403.92

Table 3: The salient characteristics of the four ways to construct the 10-fold split of the data.

Classifier	Precision	Recall	F-score
unweighted/remove			
B	3.98	79.88	7.36
S	50.81	31.69	31.94
C	82.99	49.79	54.30
weighted/remove			
B	54.96	34.22	34.14
S	60.84	46.44	49.81
C	62.69	52.88	56.47
unweighted/keep			
B	1.89	60.00	3.65
S	58.08	38.05	41.61
C	78.64	28.92	41.20
weighted/keep			
B	59.31	29.43	33.50
S	67.72	46.26	53.76
C	60.93	49.98	53.33

Table 4: Averaged precision, recall and F-score over all test partitions for each evaluation strategy. B = baseline classifier, S = classifier with path structure features, C = classifier with confidence and structure features.

the 10 folds with equal node count partitioning (see Table 2). Especially with the dense fold no. 1, the network density and therefore path count differs substantially between the training and test set, leading to the poor classification performance. With partitions balanced by edge counts, on the other hand, the baseline classifier performance is much higher, with F-scores of over 30%.

Classifiers using structure and confidence features clearly outperform the baseline in all evaluation strategies, indicating that this problem indeed is learnable and that the paths themselves, not only their overall count, provide useful information to the classification. Interestingly, the confidence features decrease the performance in evaluation strategies where paths are allowed to span across folds. As these features provide clear improvement when the folds are completely independent, further work is required to examine whether it is the case that confidence features are beneficial only with sparser networks, leading to potential gains in networks for less studied organisms.

For the more complex task of predicting also the edge type and direction we select only one evaluation strategy: balanced edge counts with edges spanning over folds. This method is chosen as it provides a sensible baseline and low variation between the folds, yet it reflects the natural density of the graph well. As the edge types are not exclusive, multiple labels can be predicted for each example, reflecting cases where several relationships exist simultaneously between the candidate GGPs, for example both Binding and Regulation.

Results for the multi-label classification task are shown in Table 5. As can be expected, the performance for this task is lower than for the simple binary classification task. As with the binary task, the performance of the classifiers is substantially higher than for the baseline. An interesting difference can be observed between the performance of predicting binding versus regulation. As binding edges are symmetric and the most common out of these types, predicting them should be intuitively the easiest. However, the baseline classifier obtains higher scores for regulation events. On the other hand, the classifier with path structure features performs clearly better for binding edges, resulting in approximately 10pp higher F-score than for regulation edges.

Indirect regulations are clearly the hardest types to predict. This might be due to their low number in the data sets or the fact that an indirect regulation edge always originates from a complex regulation event. The confidence features do not seem to have a significant influence on the results as also observed in the binary classification task. Further investigation is needed to clarify these evaluation numbers.

Edge type	Precision	Recall	F-score
B			
Binding	63.07	9.27	14.84
Reg. >	66.01	11.36	17.74
Reg. <	61.37	14.35	21.38
Ind. reg. >	36.00	5.46	9.18
Ind. reg. <	31.83	5.76	9.56
Micro-average	60.96	10.63	16.73
Macro-average	51.66	9.24	14.54
S			
Binding	66.46	37.79	46.85
Reg. >	65.14	27.80	37.47
Reg. <	64.14	27.65	36.49
Ind. reg. >	50.67	9.59	15.82
Ind. reg. <	32.94	8.99	13.87
Micro-average	65.11	30.85	40.52
Macro-average	55.87	22.36	30.10
C			
Binding	62.33	40.43	47.92
Reg. >	65.72	28.10	37.86
Reg. <	63.85	27.59	36.03
Ind. reg. >	59.38	10.74	17.73
Ind. reg. <	34.31	9.22	14.28
Micro-average	62.40	32.24	41.34
Macro-average	57.12	23.22	30.76

Table 5: Averaged precision, recall and F-score over all test partitions for each edge type. Binding is a symmetric interaction whereas regulation and indirect regulation are directed. The direction is denoted with > and <. B = baseline classifier, S = classifier with path structure features, C = classifier with confidence and structure features.

5.4 Manual Evaluation

The false positive predictions provide an extremely interesting research target from the hypothesis generation perspective. First, some of these predictions are evaluated as false positives only because the text mining system that was used to generate the underlying data has failed to extract these relationships from the text, even though they were present. And second, some of the predictions evaluated as false positives may in fact constitute existing undiscovered relationships, identification of which, after all, is the overall goal of this work.

If some proportion of the former can be found, it may at least hint at the possibility of the latter being present among the "false" positives as well. To assess whether some of the false positives can be attributed to extraction failures of the text mining system underlying the EVEX network, we manually evaluated from each partition the 10 false positive pairs with the highest number of connecting paths, 100 examples in total. This evaluation was carried out using the edge-balanced partitioning with edges spanning across partitions and the predictions were made with the classifier using the path structure features. In this evaluation we only determined whether an interaction exists between the genes and did not consider the interaction types.

The manual evaluation was carried out in two ways: First, we searched in the EVEX resource for occurrences of every false positive pair, but this time including also event occurrences among sequence homologs of the candidate genes. Given a false positive pair *geneA–geneB*, we thus inspect all pairs *geneX–geneY* such that *geneX* and *geneA* belong to one homologous family, and similarly also for *geneY* and *geneB*. This way, we are able to detect corresponding events that are reported to occur between similar genes in other organisms, instead of focusing only on the human gene regulatory network. EVEX contains several gene family definitions — we use HomoloGene (Sayers et al., 2012) and Ensembl (Flicek et al., 2013) for the evaluation, as these specifically focus on eukaryotes and include the human genome. For 15 of the 100 pairs, we found a corresponding event among HomoloGene families. Among families based on the Ensembl resource, the number of pairs was 34. Further examining these pairs, we found that 3 out of the 15 HomoloGene-based interactions (4 out of 34 with Ensembl) could be confirmed to hold among the exact human genes predicted, but the pair was not present in EVEX because of gene symbol normalization failure. These are thus cases of successful prediction of relationships not present in the EVEX network, which could be subsequently verified in the literature. The remaining 12 interactions were either reported to happen in other organisms or they were protein complex interactions and the exact subunits were not mentioned. For instance, predicted interacting genes PTK2B and NGF are found to belong to interacting families, with the sentence *"NGF induced the tyrosine phosphorylation of RAFTK..."* supporting this prediction. Even though the family assignment has grouped PTK2B together with RAFTK, the precise relation is that PTK2B is a subunit of RAFTK and no confirmed interaction is known

between PTK2B and NGF. Nonetheless, it is intriguing to observe that the system is able to predict a hypothetical interaction close to a known interaction of related protein complexes.

In the second manual evaluation, the 100 pairs were searched from the STRING database (Franceschini et al., 2013), which combines protein-protein interaction evidence from various sources, including text mining resources, experimental data and curated databases. In this evaluation all STRING evidence above the confidence value of 0.150 (i.e. the low confidence threshold on the STRING website) was considered as a possible interaction candidate. Out of the 100 pairs 31 were found to have some evidence in the STRING database.

These results indicate that the system is able to identify correct interactions not currently present in the EVEX network.

6 Conclusions and Future Work

In this paper, we have introduced a machine-learning hypothesis generation system, based on large-scale literature mining networks and supervised learning. We have shown that the problem is indeed learnable using features extracted from the network context of each candidate pair. The classification performance is far above the random baseline as well as the baseline classifier which only considers the number of paths connecting the candidate in the network. This indicates that not only the density but also the content in the network context is used by the classifier.

In addition to the aforementioned machine learning results, we have also explored some of the difficulties associated with machine learning in densely connected networks, where independence of the individual instances does not hold in many cases, causing problems in the application of the standard cross-validation procedure. Another problematic issue is the non-uniform density of the network where even few highly-connected hub nodes may cause large variance in experimental results.

There is a number of future directions for this work. First, the EVEX network offers aggregation of events not only by their Entrez Gene identifiers, but also by gene families defined through gene sequence homology and spanning across species. Incorporating events from different organisms would allow us to include the aspects of cross-species, homology based function prediction commonly used in genome annotation. Second, we currently only utilize features from the network, but not from the underlying text. It would be of interest to explore what other features from the texts, beyond the events themselves, can contribute to the classification.

Acknowledgments

We would like to thank Sofie Van Landeghem, Ghent University, for her valuable suggestions and comments, the Academy of Finland and Turku Centre for Computer Science (TUCS) for funding the study and CSC – IT Center for Science Ltd. for computational resources.

References

Tanja Bekhuis. 2006. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, 3(1):2.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at PubMed scale. *Bioinformatics (ISMB'2010 proceedings volume)*, 26:i382–i390.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13(Suppl 11):S4.

Paul Flicek, Ikhlak Ahmed, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Carlos Garca-Girn, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Thomas Juettemann, Andreas K. Khri, Stephen Keenan, Monika Komorowska, Eugene Kulesha, Ian Longden, Thomas Maurel, William M. McLaren, Matthieu Muffato, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet Singh Riat, Graham R. S. Ritchie, Magali Ruffier, Michael Schuster, Daniel Sheppard, Daniel Sobral, Kieron Taylor, Anja Thormann, Stephen Trevanion, Simon White, Steven P. Wilder, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Jennifer Harrow, Javier Herrero, Tim J. P. Hubbard, Nathan Johnson, Rhoda Kinsella, Anne Parker, Giulietta Spudich, Andy Yates, Amonida Zadissa, and Stephen M. J. Searle. 2013. Ensembl 2013. *Nucleic Acids Research*, 41(D1):D48–D55.

Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars Juhl Jensen. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, 41(Database-Issue):808–815.

Martin Gerner, Farzaneh Sarafraz, Casey M Bergman, and Goran Nenadic. 2012. BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, 28(16):2154–2161.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*.

Jin-Dong 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.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA, June. Association for Computational Linguistics.

Dominique Lasalle and George Karypis. 2013. Multi-threaded graph partitioning. *Parallel and Distributed Processing Symposium, International*, 0:225–236.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria, August. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC bioinformatics*, 9(Suppl 3):S6.

Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Evan Bolton, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M. Fingerman, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, Sergey Krasnov, David Landsman, David J. Lipman, Zhiyong Lu, Thomas L. Madden, Tom Madej, Donna R. Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Karsch-Mizrachi, James Ostell, Anna Panchenko, Lon Phan, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Yanli Wang, W. John Wilbur, Eugene Yaschenko, and Jian Ye. 2012. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 40(D1):D13–D25.

Don R. Swanson. 1988. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557.

Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6(7):e1000837, 07.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE*, 8(4):e55814.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2012. SR4GN: A species recognition software tool for gene normalization. *PLoS ONE*, 7(6):e38460, 06.

Incorporating Topic Modeling Features For Clinic Concept Assertion Classification

Dingcheng Li[1], Ning Xia[2], Sunghwan Sohn[1], Kevin B. Cohen[3], Christopher G. Chute[1], Hongfang Liu[1]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 55901, USA

[2]Case Western Reserve University, Cleveland, Ohio, 44106, USA

Biomedical Text Mining Group, Computational Bioscience Program, University of Colorado Health Science Center

{li.dingcheng, sohn.sunghwan, chute, liu.hongfang}@mayo.edu, ning.xia@case.edu, kevin.cohen@uchsc.edu

Abstract

With the rapid growth of electronic medical records (EMR), clinical information resides over abundantly in clinical narratives. Natural language processing (NLP) techniques have been applied to unlock such information with promising results. One of the popular NLP techniques used is supervised text classification. In supervised text classification, a collection of instances (e.g., documents or sentences), annotated with labels, can be used to train a classifier to assign labels to an un-labeled instance. Traditionally, features used in text classification are unigrams, bigrams, and/or concepts. In this study, we explore the use of Latent Dirichlet Allocation (LDA) to generate topic features to capture latent semantics and incorporate them into text classification. We applied the method on the well-known clinical concept assertion task in the i2b2 2010 NLP challenge. The result shows significant improvement when incorporating topic modeling features into text classification (by 3.69 percent). Additionally, the inferred topic distribution offers the latent semantic interpretation.

1 Introduction

With the rapid growth of electronic medical records (EMRs), it becomes more and more essential to develop methods to automatically extract clinical information from EMRs, including medical entities, relations between entities, and their corresponding attributes in a timely and accurate manner. Multiple natural language processing (NLP) techniques have been applied to unlock such information from clinical narratives. Most NLP applications use pattern-based information extraction (IE) [1, 2] where lexical, syntactic and semantic patterns are manually engineered for extracting related information from text. Recently, supervised text classification has become popular where a collection of labeled instances (e.g., documents or sentences or named entities) can be used to train a classifier to assign labels to an un-labeled instance [3, 4]. Common features for text classification are surface level features including unigrams, bigrams, and/or concepts. A diverse range of classification algorithms such as support vector machines (SVMs) [5], neural network, conditional random fields (CRFs) [6] have been used.

Rcently, Latent Dirichlet Allocations (LDA) [7] has gained popularity in diverse fields due to the fact that it holds great promise as a means of gleaning actionable insight from the text or image datasets. LDA clusters both words and documents into topics by approximating word or term distributions.

In this paper, we explore the incorporation of topic features acquired using LDA into text classification in the domain under the assumption that a mixture of common topics can be

discovered from a collection of clinical documents and those topics can be discriminative for text classification.

2 Background and Related Work

In the following, we present the background information of LDA and summarize related work of utilizing topic modeling in biomedical informatics.

2.1 LDA

In text analysis, LDA represents a document as a mixture of fixed topics, each topic *z* has the weight $\boldsymbol{\theta_z^s}$ in passage s and each topic is a distribution over a finite vocabulary of words, and each word *w* has a probability $\boldsymbol{\phi}$ in topic *z*. Placing symmetric Dirichlet priors on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, with $\boldsymbol{\theta \sim Dirichlet(\alpha)}$ and $\boldsymbol{\phi^z \sim Dirichlet(\beta)}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are hyper-parameters to control the sparsity of distributions, the generative model is given by:

$$w_i|z_i, \phi_{w_i}^{z_i} \sim Discrete(\phi^{z_i}), i = 1, \dots, W \quad \textbf{Eq-1}$$

$$\phi^z \sim Dirichlet(\beta), z = 1, \dots, K \quad \textbf{Eq-2}$$

$$z_i|\theta^{s_i} \sim Discrete(\theta^{s_i}), i = 1, \dots, W \quad \textbf{Eq-3}$$

$$\theta^s \sim Dirichlet(\alpha), s = 1, \dots, S \quad \textbf{Eq-4}$$

where *K* is the total number of topics, W is the total number of words in the document collection, and $\boldsymbol{s_i}$ and $\boldsymbol{z_i}$ are the passage and the topic of the *i*th word $\boldsymbol{w_i}$ respectively. Each word in the vocabulary $\boldsymbol{w_i \in V = [w_1, w_2, \dots, w_W]}$ is assigned to each latent topic variable $\boldsymbol{z_i}$. Given a topic $\boldsymbol{z_i = k}$, the expected posterior probability $\boldsymbol{\hat{\theta}^s}$ of topic mixings of a given passage s and the expected posterior probabilities $\boldsymbol{\hat{\phi}_{w_i}^{z_i}}$ of a word $\boldsymbol{w_i}$ are calculated as below.

$$\hat{\phi}_{w_i}^{z_i} = \frac{n_{w_i}^k + \beta}{\sum_{j=1}^{W} n_{w_j}^k + W\beta} \quad \textbf{Eq-5}$$

$$\hat{\theta}^s = \frac{n_s^k + \alpha}{\sum_{j=1}^{K} n_s^j + K\alpha} \quad \textbf{Eq-6}$$

where $n_{w_i}^k$ is the count of $\boldsymbol{w_i}$ in topic *k*, and n_s^k is the count of topic *k* in passage *s*.

2.2 Topic Modeling in Biomedical Informatics

In biomedical informatics, probabilistic topic modeling has been applied to patients' notes to discover relevant clinical concepts and relations between patients [8]. Angues et al. [9] applied unsupervised LDA to primary clinical dialogues for visualizing shared content in communication. Wang et al. developed BioLDA [10] to find complex biological relationships in recent PubMed articles. Wu and Xu [11] made use of LDA to rank gene-drug relationships in biomedical literatures based on Kullback-Leibler (KL) distance between topics derived from LDA. Bisgin et al. [12, 13] mined FDA drug labels using topic modeling. Fifty-two unique topics, each containing a set of terms, were identified and then the probabilistic topic associations were used to measure the similarity between drugs. Zhou et al. [14] utilized the topic features to categorize the collections tweets into latent topics and those topics are used as features to train SVM prediction models for mining adverse effects labels. Newman et al [15] and Bundachus and Tresp [16] employed topic models to interpret MeSH terms. Chen et al. [17] proposed to use LDA to promote ranking diversity for genomics information retrieval and they claimed that topic distributions of retrieval passages can help identify aspects more accurately. Chen et al. [18] extended LDA by including background distribution to study microbial samples. Under their setting, each microbial sample is a document and each functional element is a word. They found that estimating the probabilistic topic model can uncover the configuration of functional groups. All of those studies have shown the potentiality of topic modeling.

3 Experimental Methods

In this study, our main goal is to investigate the effectiveness of topic modeling in facilitating text classification. We first generate topic distribution for a collection of documents and then incorporate the generated topics as features for text classification. In the following, we provide the description of the text classification task**.**

3.1 Clinic Concepts Assertion

We use the clinical concept assertion data set used in the i2b2 2010 NLP challenge. The data set consists of 11960 concepts from 394 reports in the training set and 18850 concepts from 477 reports in the test set. The class label for each concept refers to the status of how the medical concept pertains to the patient. There are six class labels: *present*, *absent*, *hypothetical, possible*, *conditional* and *association with someone else* (*AWSE* for short). An assertion classifier can be trained to assign a class label to a concept based on its corresponding context text. For example, in the sentence, "*did not have an oxygen requirement upon discharge*", "*an oxygen requirement*" is the concept with concept type as "*problem*" and assertion class label as "*absent*". An assertion classifier is supposed to assign one of the six class labels to a problem concept. For this example, if the label "*absent*" is assigned to the concept "*an oxygen requirement*", it implies that the classifier classifies the status of the concept or the problem correctly.

3.2 Topics Generated as Features for Classification Models

In our setting, words surrounding the assertion concepts are set as the document. The number of topics experimented ranges from 6 to 102 with 6 as the incremental number. We use the package JGibbsLDA [19] for topic model generation. This package implements Gibbs sampling to estimate parameters θ and ϕ. Gibbs sampling is a special case of Markov-chain Monte Carlo (MCMC) [20] and often yields relatively simple algorithms than Variational Methods [21] for approximating inferences in high-dimensional models including LDA [22] though usually larger number of iterations is needed. In the I2B2 corpus, one or two assertion concepts are annotated within one sentence. Hence, the "document" in our work is short, only about 20 or fewer words involved. In this case, 200 iterations for the Gibbs sampler are sufficient enough to yield a good estimation of the parameters. Then, the posterior $\hat{\theta}^{s}$ for documents and $\hat{\phi}_{w_i}^{z_i}$ for words are employed as topic features to train classifiers.

3.3 Classifier selection

We use LibSVM [23], one of the most popular implementations for SVM, for classification. For SVM, firstly, we need to determine the kernel function and corresponding parameter settings. In previous work, researchers claim that the linear function with *C* as *20000* achieves the best results on rich set of features [24]. Nonetheless, they do not mention how and why they select *20000* as the value of *C*. We test linear function with the same setting. Yet, we find that model trained with radial basis function (RBF) with suitable setting of *C* and *gamma* (*C*=32 and *gamma*=0.0079) achieve the best results. Those settings are found with grid search strategy. There are two advantages using RBF kernel over using linear one. With grid search on RBF, we can find much lower *C*, and thus the training speed is much faster and secondly, the parameters chosen with grid search strategy are more empirically sound.

3.4 Feature extraction

We adopt common features used by most i2b2 NLP challenge participants as the base line. The following summarizes them:

BOW, bigrams and concepts - Those features include the concept term itself, the four words preceding it, and the four words following it. We use the LVG annotator in Lexical Tools [25] to normalize each word (e.g., with respect to case and tense). Meanwhile, for the four words on the left and right of the concepts, bigrams are also composed.

Contextual Features - We incorporate the ConText algorithm [26] to detect four contextual properties in the sentence: absent (negation), hypothetical, historical, and not associated with the patient. The algorithm assigns one of three values to each feature: true, false,

or possible. Furthermore, it is found that locations of those contextual phrases can be discriminative. Hence, we also distinguish whether the contextual feature is before and after the concept.

Section features - Clinic notes are usually composed of two sections: admission and discharge. They are quite related to patient status and a feature easily extracted. We create one feature to represent the Section Header with a string value normalized using LVG again.

Orthographical features – We also define orthographical features to capture affixes, capitalization, mixture of digits and letters, all digits and so on.

Topic features – We incorporate the posterior distributions of the document against the topic and those of the words against the topic generated from unsupervised LDA.

Table 1 shows the example features generated for the assertion task of "an oxygen requirement" in sentence "*did not have an oxygen requirement upon discharge"*.

Table 1. Features extracted for the example sentence.

Feature Types	Features extracted
BOW with window 4	*did, not, have, upon, discharge,*
Bigrams	*did-not, not-have, upon-discharge*
Concepts	*an, oxygen, requirement*
Contextual Features	not, discharge
Section features	admission (found from the original text)
Orthographical Features	none in this example
Topic Features	topic 1: 0.05, topic 2: 0.25, …

3.5 Experimental setting

Three groups of experiments are set up, namely, baseline (including all features except topic features), topic features as numerical values, and topic features as Boolean values. The baseline features are represented as bag of word (BOW) and the values are simply Boolean. Unlike other features, topic distributions are in numerical forms ranging from 0 to 1. Therefore, for the second group, we mixed the numerical values from topic features and other Boolean features. In the third one, we convert the numerical representation to Boolean values as well so that all features have consistent representations. Some threshold is selected for topic numerical values, higher than the threshold as 1, otherwise as 0.

3.6 Evaluation metrics

We use standard evaluation metrics, namely, precision, recall and F-measure to evaluate the performance. We report both micro and macro metrics, where the former one computes the overall performance and the latter as the average across each class.

4 Results and discussion

4.1 Generating Topic Features

In this work, we tested different topic numbers when generating topic features in order to examine how Gibbs Sampling influences the classification accuracy. Topic numbers examined ranges from 6 to 102 with an increment step as 6. We assumed symmetric priors and set hyper-parameters following the principle defined in Jun Zhu [27]. In the assertion classification task, there are 11,100 words in the training data and 10,934 words in the testing data after stop words were removed using the list provided by the Mallet package [28] and words were normalized with the porter stemmer [29]. The optimal result was obtained when the topic number was set to 12.

Table 2. Evaluation experimental results.

Feature Set	**Precision Micro (macro)**	**Recall Micro (Macro)**	**F-Measure Micro (Macro)**
Baseline features as Boolean values	0.8931 (0.5908)	0.8914 (0.7770)	0.8922 (0.6713)
Baseline + Numerical Topic features	0.9129 (0.6327)	0.9129 (0.8378)	0.9129 (0.7209)
Baseline + Boolean Topic features	0.9250 (0.6913)	0.9285 (0.9046)	0.9268 (0.7837)

4.2 Classification results

Table 3. Confusion matrix for all three experiments

	Baseline (Baseline + Numerical topic features) Baseline + Boolean topic features					
truth / predication	**absent**	**AWSE**	**conditional**	**hypothetical**	**possible**	**present**
absent	**3113 (3260) 3326**	30 (17) 26	7 (9) 8	35 (20) 18	33 (32) 23	191 (294) 123
AWSE	30 (7) 13	**55 (63) 107**	3 (11) 3	23 (25) 7	15 (25) 10	74 (22) 2
conditional	27 (35) 10	10 (5) 8	**31 (33) 38**	23 (22) 6	16 (22) 21	72 (22) 3
hypothetical	30 (38) 40	8 (19) 7	3 (1) 2	**513 (579) 590**	25 (47) 9	118 (180) 50
possible	40 (53) 24	12 (13) 9	3 (3) 4	33 (22) 20	**390 (374) 410**	144 (180) 75
present	244 (205) 228	30 (134) 36	124 (125) 140	90 (76) 76	404 (383) 401	**12434 (12611) 12752**

It was found from the micro evaluation that with baseline features, the prediction micro F-measure was as high as 89% and macro F-measure as 67%, consistent with what was achieved in previous studies [30]. After topic features were incorporated, the micro F-measure increased from 0.8922 to 0.9129 for numerical topic features and 0.9268 after numerical topic features were converted to Boolean topic features using the threshold of 0.1 (the threshold is determined by heuristic methods, namely, with simple observation of the distribution of topic values and a few trials). For macro metrics, there were about 0.0419 and 0.0608 increase in precision and recall respectively when using topic features as numerical features and 0.0586 and 0.0668 increase when using topic features as Boolean values.

4.3 Confusion matrix

Table 3 shows the confusion matrix for all six assertion class labels which can provide some insights on the contribution of topic features in assertion classification. From Table 3, we can see that when incorporating topic features as numerical values, the true positive rate for almost all assertion classes has increased except *possible* class with a decrease of 16, all assigned to *hypothetical*. It implies that *possible* and *hypothetical* tend to mix together. When topic features are incorporated as Boolean values, the performance becomes even better with an increase for all assertion classes.

5 Topic distribution analysis

In this section, we performed in depth analysis of topic distributions.

5.1 Likelihood estimation

In LDA, estimation of topic distributions of words was evaluated with Log-likelihood score of the posterior distribution of words given topics, one of the standard criteria for generative model evaluation. It provides a quantitative measurement of how well a topic model fits the training data. It is defined as the integrating out of all latent variables shown in Eq-7. The higher the score, the better the model fitness.

In the Gibbs sampling iterations, the equation is simplified as the product of $\hat{\phi}_{w_i}^{z_i}$ and $\hat{\theta}^s$. In **Figure 1**, it shows the trend that the log-likelihood goes with the change of the number of topics when we run LDA. When topic numbers increase from 6 to 102, the log-likelihood firstly decreases and after the number reaches around 50, the log-likelihood starts to increase. This is somewhat different from the common trend found in other domains [18]. Usually, the log-likelihood would increase with the increase of the topic number and then after some peak, it would decrease. The peak is usually selected as the best number of topic. Probably, the trend observed in this corpus is related to

the text classification task defined here where we used short sentences as documents.

$$p(\boldsymbol{w}|\boldsymbol{z}) = \prod_{t=1}^{T}\left[\int_{\phi_{z_t}} p(\boldsymbol{w}|z_t, \phi_{z_t})p(\phi_{z_t})d\phi_{z_t}\right] \qquad \textbf{Eq-7}$$

$$= \left[\frac{\Gamma(W\mathrm{B})}{\Gamma(\mathrm{B}^W)}\right]^T \frac{\prod_{\mathrm{w_i}}\Gamma(n_t^{w_t}+\alpha)}{\Gamma(n_t^{(.)}+W\alpha)}\left[\frac{\Gamma(W\mathrm{B})}{\Gamma(\mathrm{B}^W)}\right]^T \prod_{t=1}^{T}\frac{\prod_{\mathrm{w_i}}\Gamma(n_t^{w_t}+W\mathrm{B})}{\Gamma\left(n_t^{(.)}+W\mathrm{B}\right)}$$

5.2 Illustration of connections between topic assignment of word and the assertion classes

Since when topic number is 12, the classifier achieves the best result, we used the corresponding topic distribution to illustrate the connections between the topic assignments and the six assertion classes.

Figure 2 is the graph to visualize the proportion of each topic associated with each assertion class. As shown here, almost each class has its own dominant topics. There are about 0.27 of *absent* falling into the second topic and about 0.16 into the third topic. *AWSE* spreads from topic 3, 7, 8 and 9. *Conditional* involves topic 4, 7 and 3 while *hypothetical* topic 6, 3, 4 and 7. The highest one for *possible* is the 6th topic. Similarly, *present* falls into 10th, 5th, 7th and 3rd.

Figure 3 shows the word cloud generated for each topic. The figure was generated with wordle [31]. We selected top 50 words generated by LDA for each of the 12 topics and generated sample documents based on word probability assigned by the LDA training model as the input to wordle. The font size of the word shows how important the word in that topic. If we compared Figure 2 and Figure 3, there are some clear alignments between the assertion class and topics. The following describes our investigation in detail.

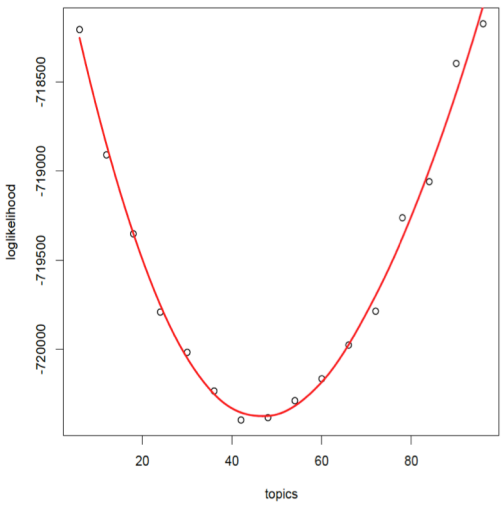


Figure 1. **Log-Likelihood distribution trend**

Absent - The dominant topic for class *absent* is topic 2. In the word cloud, we see *without*, *no* or *deny* are the majority words in topic 2. In addition, we can find that topic 11, though not much proportion in *absent* either, is much more in *absent* than in other classes. Terms as *fever*, *nausea*, *bleed* and *infection* can be seen in it besides *deny*. It may indicate that those are popular signs and symptoms physicians tend to record its presence or absence.

AWSE - When we move on to the second class, AWSE, it is obvious to see that topic 9 is the largest component. Words like *family* or *history*, which is outstanding in Topic 9 tells that it is related to AWSE. *Conditional and hypothesis* - Topic 4 is mostly in *conditional* and *hypothesis*, in particular, in *conditional*. Firstly, terms like *afebrile*, *wind*, *problem*, *tumor*, *issue* or *episode* are often used to show some potential symptoms. Verbs, like *become*, *resolve*, *control* or *remain*, adjectives or connectives, like *dry* or *far* or *if* are often related to predictions or deductions.

Hypothetical and possible - Classes *hypothetical* and *possible* share more even portions of topics than others, especially topic 6 and topic 7, since both of them include *if* and diverse disease terms. After we browsed the corpus, we found that those disease names are usually referring acute ones and usually happen under some conditions so they generally appear as *possible* as well as *hypothetical*.

Present - The assertion class *present* is the most dominant class with 13000 concepts in total. This seems to explain why there is more even distribution for *present* than other classes. But we can still find some topics are more dominant than others. Topic 5 can be regarded to the most important one. In word cloud, *present*, *diabetes*, *mellitus chronic* and *syndrome* are seen there. Those disease terms are chronic diseases. Thus, the patient status should be present usually.

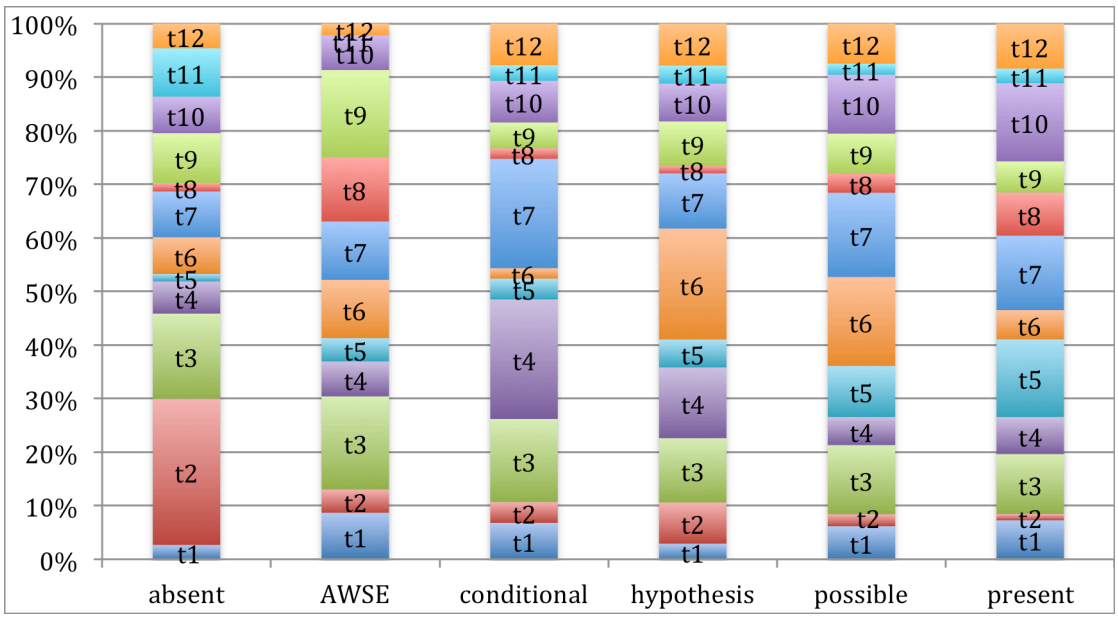


Figure 3 Proportion of topics for each assertion class where *t* stands for topic

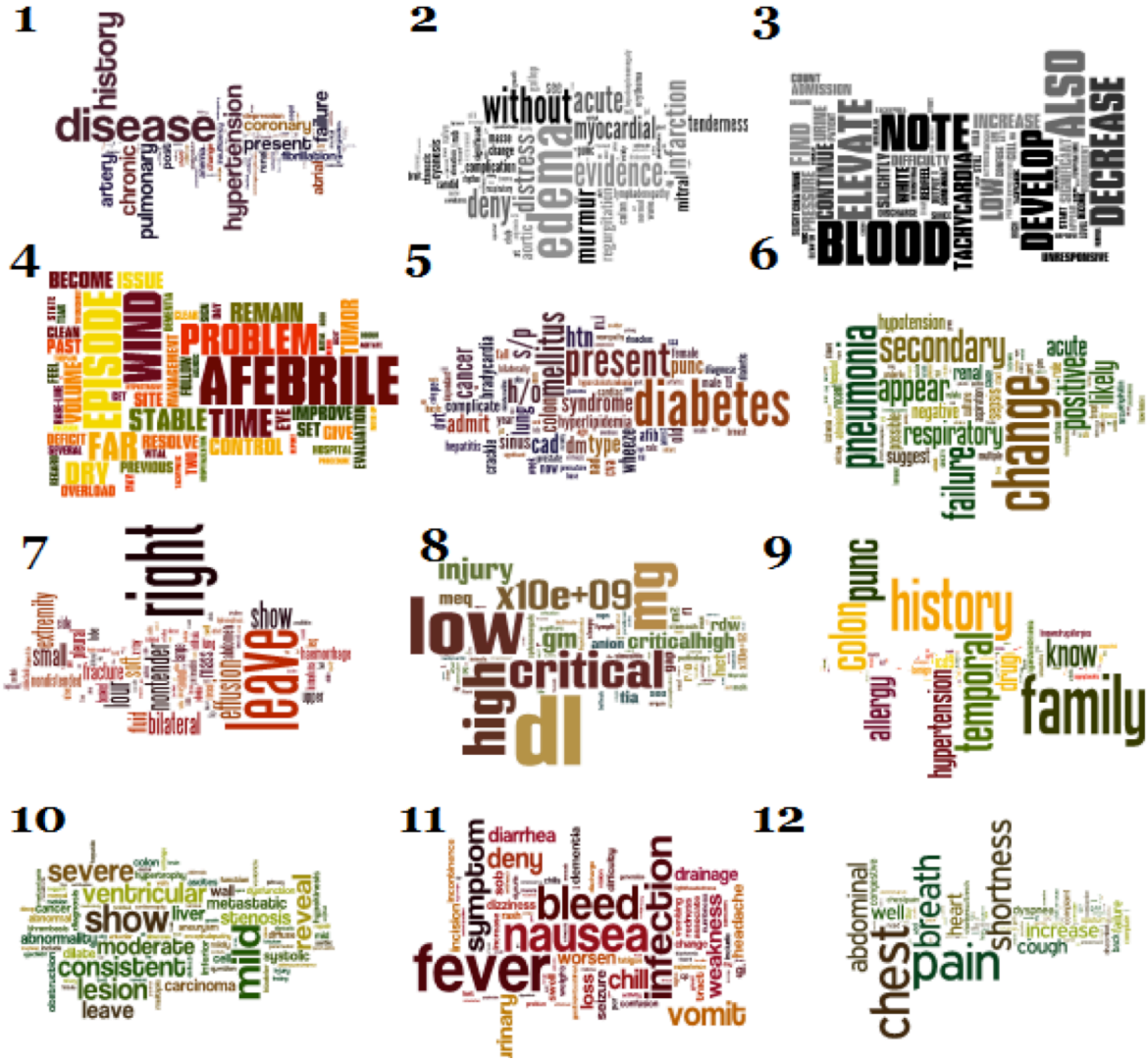


Figure 2. **Word cloud for 12 topics.**

Topic 3 and 10 also spread across all assertion classes. Words like *blood*, *develop*, *note*, *decrease* and *elevate* in topic 3*, mild, consistent, moderate, severe, liver, ventricular and show* in topic 10 and so on are big words there. It indicates they are popular terms physicians or nurses use in their work. The reason that they two are split into two topics seem to be related to the focus of the two topics. Topic 3 looks more related to blood or something changeable in quantity while topic 10 seems to be closer to body parts. Topic 8 is an interesting topic. It mainly appears in *AWSE* and *present* and involves more words like dosage or measures. This seems to suggest that it is more related to existing diseases no matter the patient himself or someone else. Topic 12, which includes words, like *chest*, *pain*, *breath, couth* and *shortness* is more related to respiratory diseases. From Figure 2, we can see all classes except *absent* and *AWSE* have similar portion about it. This seems to suggest that many patients in the corpus have respiratory diseases, but not their family members.

6 Conclusion and Future Work

The experimental results show that topic modeling enjoys its natural advantages as shown in other domains and improves the performance of text classification. However, as we discussed in previous sections, even the topics generated look discriminative but quite a few topics did not align well with assertion classes. In the future, we plan to incorporate supervision into topic modeling [32, 33] to generate topics that are discriminative for the classification task.

Acknowledgments

This study was made possible by National Science Foundation ABI:0845523, National Institute of Health R01LM009959A1 and R01GM102283A1.

References

1. Sohn, S., et al., *Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification.* Journal of the American Medical Informatics Association, 2013.
2. Savova, G., et al. *Towards temporal relation discovery from the clinical narrative*. in *AMIA Annual Symposium Proceedings*. 2009. American Medical Informatics Association.
3. Geng, B., et al., *Ensemble manifold regularization.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2012. **34**(6): p. 1227-1233.
4. Leaman, R. and G. Gonzalez. *BANNER: an executable survey of advances in biomedical named entity recognition*. in *Pacific Symposium on Biocomputing*. 2008.
5. Li, D., K. Kipper-Schuler, and G. Savova. *Conditional random fields and support vector machines for disorder named entity recognition in clinical texts*. in *Proceedings of the workshop on current trends in biomedical natural language processing*. 2008. Association for Computational Linguistics.
6. Settles, B. *Biomedical named entity recognition using conditional random fields and rich feature sets*. in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*. 2004. Association for Computational Linguistics.
7. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent Dirichlet allocation.* Journal of Machine Learning Research, 2003. **3**: p. 993-1022.
8. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic acids research, 2004. **32**(suppl 1): p. D493-D496.
9. Hersh, W.R., et al. *TREC 2006 Genomics Track Overview*. in *TREC*. 2006.
10. Wang, H., et al., *Finding complex biological relationships in recent PubMed articles using Bio-LDA.* PLoS One, 2011. **6**(3): p. e17243.
11. Wu, Y., et al. *Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation*. in *Pacific Symposium on Biocomputing*. 2012. World Scientific.
12. Bisgin, H., et al., *Mining FDA drug labels using an unsupervised learning technique-topic modeling.* BMC bioinformatics, 2011. **12**(Suppl 10): p. S11.

13. Bisgin, H., et al., *Investigating drug repositioning opportunities in FDA drug labels through topic modeling.* BMC bioinformatics, 2012. **13**(Suppl 15): p. S6.
14. Zhou, D., et al. *Exploring social annotations for information retrieval*. in *Proceedings of the 17th international conference on World Wide Web*. 2008. ACM.
15. Newman, D., S. Karimi, and L. Cavedon, *Using topic models to interpret MEDLINE's medical subject headings*, in *AI 2009: Advances in Artificial Intelligence*. 2009, Springer. p. 270-279.
16. Bimboim, H. and J. Doly, *A rapid alkaline extraction procedure for screening recombinant plasmid DNA.* Nucleic acids research, 1979. **7**(6): p. 1513-1523.
17. Chen, Y., et al., *A LDA-based approach to promoting ranking diversity for genomics information retrieval.* BMC genomics, 2012. **13**(Suppl 3): p. S2.
18. Chen, X., et al. *Inferring functional groups from microbial gene catalogue with probabilistic topic models*. in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*. 2011. IEEE.
19. Phan, X.-H., L.-M. Nguyen, and S. Horiguchi. *Learning to classify short and sparse text & web with hidden topics from large-scale data collections*. in *Proceedings of the 17th international conference on World Wide Web*. 2008. ACM.
20. Geman, S. and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1984(6): p. 721-741.
21. Griffiths, T.L. and M. Steyvers, *Finding scientific topics.* Proceedings of the National academy of Sciences of the United States of America, 2004. **101**(Suppl 1): p. 5228-5235.
22. Heinrich, G., *Parameter estimation for text analysis.* Web: http://www. arbylon. net/publications/text-est. pdf, 2005.
23. Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.
24. de Bruijn, B., et al., *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, 2011. **18**(5): p. 557-562.
25. Stede, M., *Lexical semantics and knowledge representation in multilingual text generation*. 1999: MIT Press.
26. Chapman, W.W., D. Chu, and J.N. Dowling. *ConText: An algorithm for identifying contextual features from clinical text*. in *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. 2007. Association for Computational Linguistics.
27. Zhu, J., A. Ahmed, and E.P. Xing, *Medlda: maximum margin supervised topic models.* Journal of Machine Learning Research, 2012. **13**: p. 2237-2278.
28. McCallum, A.K. *Mallet: A machine learning for language toolkit*. 2002.
29. Porter, M.F., *Snowball: A language for stemming algorithms*, 2001.
30. Uzuner, Ö., et al., *2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.* Journal of the American Medical Informatics Association, 2011. **18**(5): p. 552-556.
31. Feinberg, J., *Wordle.* Ch, 2009. **3**: p. 37-58.
32. Li, D., S. Somasundaran, and A. Chakraborty. *A combination of topic models with max-margin learning for relation detection*. in *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*. 2011. Association for Computational Linguistics.
33. Li, D., S. Somasundaran, and A. Chakraborty, *ERD-MedLDA: Entity relation detection using supervised topic models with maximum margin learning.* Natural Language Engineering, 2012. **18**(2): p. 263.

Distributional Semantics Resources for Biomedical Text Processing

Sampo Pyysalo[1] Filip Ginter[2] Hans Moen[3] Tapio Salakoski[2] Sophia Ananiadou[1]
1. National Centre for Text Mining and School of Computer Science
University of Manchester, UK
2. Department of Information Technology
University of Turku, Finland
3. Department of Computer and Information Science
Norwegian University of Science and Technology, Norway
sampo@pyysalo.net ginter@cs.utu.fi hans.moen@idi.ntnu.no
tapio.salakoski@utu.fi sophia.ananiadou@manchester.ac.uk

Abstract

The openly available biomedical literature contains over 5 billion words in publication abstracts and full texts. Recent advances in unsupervised language processing methods have made it possible to make use of such large unannotated corpora for building statistical language models and inducing high quality vector space representations, which are, in turn, of utility in many tasks such as text classification, named entity recognition and query expansion. In this study, we introduce the first set of such language resources created from analysis of the entire available biomedical literature, including a dataset of all 1- to 5-grams and their probabilities in these texts and new models of word semantics. We discuss the opportunities created by these resources and demonstrate their application. All resources introduced in this study are available under open licenses at http://bio.nlplab.org.

1 Introduction

Despite efforts to create annotated resources for various biomedical natural language processing (NLP) tasks, the number of unannotated domain documents dwarfs that of annotated documents by many orders of magnitude. The PubMed literature database provides access to over 23 million citations, of which nearly 14 million include an abstract. The biomedical sciences are also at the forefront of the shift toward open-access (OA) publication (Laakso and Björk, 2012), with the PubMed Central (PMC) OA subset containing nearly 700,000 full-text articles in an XML format.[1] Together, these two resources constitute an unannotated corpus of 5.5 billion tokens, effectively covering the entire available biomedical scientific literature and forming a representative corpus of the domain (Verspoor et al., 2009).

The many opportunities created by the availability of large unannotated corpora for various NLP methods are well established (see e.g. Ratinov and Roth (2009)), and models induced from unannotated texts have been considered also in a number of recent biomedical NLP studies (Stenetorp et al., 2012; Henriksson et al., 2012). A particular focus of recent research interest are models of meaning induced from unannotated text, with numerous methods introduced for capturing both the semantics of words as well as those of phrases or whole sentences (Mnih and Hinton, 2008; Collobert and Weston, 2008; Turian et al., 2010; Huang et al., 2012; Socher et al., 2012). Although such approaches generally produce better models with more data, their computational complexity has largely limited their application to corpus sizes far below that of the biomedical literature. Recently, a number of efforts have introduced new language resources derived from very large corpora and demonstrated approaches that allow word representations to be induced from corpora of billions of words (Lin et al., 2010; Mikolov et al., 2013). However, despite the relevance of such approaches to biomedical language processing, there have to the best of our knowledge been no attempts to apply them specifically to the biomedical literature.

Corpora containing billions of words can represent challenges even for fully automatic processing, and most domain efforts consequently focus

[1] In this study, we do not consider PDF supplementary materials (see e.g. Yepes and Verspoor (2013)).

	Subset		
	PubMed	PMC OA	Total
Documents	22,120,269	672,589	22,792,858
Sentences	124,615,674	105,194,341	229,810,015
Tokens	2,896,348,481	2,591,137,744	5,487,486,225

Table 1: PubMed and the PMC OA statistics, representing the entire openly available biomedical literature. Note that PubMed statistics omit documents found also in PMC OA, and that only approximately 14 million of PubMed documents include an abstract.

n	#
1	24,181,640
2	230,948,599
3	1,033,760,199
4	2,313,675,095
5	3,375,741,685

Table 2: Counts of unique n-grams.

AFUB_038070
epicardin/capsulin/Pod-1-mediated
22-methoxydocosan-1-ol
mmHg/101.50+/-12.86
5.26@1000
40.87degrees
electromyocinesigraphic
(1-5)-KDO
overpressurizing
rootsanel

Table 3: A random sample of 10 tokens appearing exactly once in the openly available literature.

only on small subsets of the literature at a time. To avoid duplication of efforts, it is therefore desirable to build and distribute standard datasets that can be utilized by the community. In this work, we introduce and evaluate new language resources derived from the entire openly available biomedical scientific literature, releasing these resources to the community under open licenses to encourage further exploration and applications of literature-scale resources for biomedical text processing.

2 Materials and methods

2.1 Text sources

Article titles and abstracts were drawn from the PubMed distribution as of the end of September 2013, constituting in total 22,723,471 records. Full-text articles were, in turn, sourced from the PubMed Central Open Access (PMC OA) section, again as of the end of September 2013, and constitute 672,589 articles. PubMed abstracts for articles that are also present in PMC OA were discarded, so as to avoid the duplication of the abstract, which is also part of the PMC full text.

2.2 Text preprocessing

We first extracted document titles and abstracts from the PubMed XML and extracted all text content of the PMC OA articles using the full-text article extraction pipeline[2] introduced for the BioNLP Shared Task 2011 (Stenetorp et al., 2011). Since the pipeline extracts all text content, also including sections not desired for the current resource such as author affiliations and lists of references, we used a custom script to post-process the output and preserve only text from the title, abstract, and main body of the articles. We further removed inline formulae. Both for the abstracts and the full-text articles, Unicode characters were mapped to ASCII using the replacement table also used in the BioNLP Shared Task pipeline. This step is motivated by the number of commonly used NLP tools which do not handle Unicode-encoded text correctly, as well as the normalization gained from mapping, for example, the character β to the ASCII string *beta* — both of which are common in the input text. The extracted text was then segmented into sentences using the GENIA sentence splitter[3] and tokenized using a custom tokenization script replicating the tokenizer used in the GENIA Tagger (Tsuruoka et al., 2005). The resulting corpus consists in total of 5.5B tokens in 230M sentences. Detailed statistics are shown in Table 1.

2.3 N-grams

All 1- to 5-grams from the data were collected using the KenLM Language Model Toolkit (Heafield et al., 2013) and a custom tool[4] based on HAT-tries (Askitis and Sinha, 2007). The counts of unique

[2] https://github.com/spyysalo/nxml2txt

[3] https://github.com/ninjin/geniass

[4] https://github.com/spyysalo/ngramcount

Word2vec				Random Indexing			
Input: cysteine		Input: methylation		Input: cysteine		Input: methylation	
Word	Distance	Word	Distance	Word	Distance	Word	Distance
cystein	0.865653	hypermethylation	0.815192	lysine	0.975116	hypermethylation	0.968435
serine	0.804936	hypomethylation	0.810420	proline	0.968552	acetylation	0.967535
Cys	0.798540	demethylation	0.780071	threonine	0.963178	fragmentation	0.961802
histidine	0.782239	methylated	0.749713	arginine	0.963163	plasticity	0.960208
proline	0.771344	Methylation	0.749538	histidine	0.962816	hypomethylation	0.959995
Cysteine	0.769645	methylations	0.745969	glycine	0.960027	replication	0.959925
aspartic	0.750118	acetylation	0.740044	tryptophan	0.959929	deletions	0.956500
active-site	0.745223	DNA-methylation	0.739505	methionine	0.959649	disturbance	0.955987
asparagine	0.735614	island1	0.738123	serine	0.958578	pathology	0.954187
cysteines	0.725626	hyper-methylation	0.730208	Cys	0.953123	asymmetry	0.953079

Table 4: Nearest words for selected inputs in the two models.

n-grams are shown in Table 2. Of the 24M unique tokens, a full 14M are singleton occurrences. To illustrate the long tail, ten randomly selected singleton tokens are shown in Table 3.

Having precomputed all n-grams enables an efficient way of building word vectors, utilizing the fact that the list of n-grams includes all unique windows focused on each word in the corpus together with their count (or, correspondingly, probability). This makes the n-gram model a compressed representation of the corpus with all salient information needed to build a distributional similarity model. As opposed to the standard technique of sliding a window across the corpus, one can instead aggregate the information directly from the n-grams.

2.4 Word vectors from n-grams with Random Indexing

Random indexing (Kanerva et al., 2000) is a method for building a semantic word vector model in an incremental fashion. First, every word is assigned an *index vector* with all elements equal to zero, except for a small number of randomly distributed +1 and -1 values. The vector space representation of a given word is then obtained by summing up the index vectors of all words in all its context windows in the corpus.

We used an existing implementation of random indexing[5] that we modified to consider each 3-gram as the left half window of the rightmost word, as well as the right half window of the leftmost word. The index vectors are weighted by their corresponding probability. For the training we used vector dimensionality of 400, 4 non-zeros in the index vectors, and shifted index vectors in the same way as was done for *direction vectors* by Sahlgren et al. (2008). We also weighted the index vectors by their distance to the target word according to the following equation: $weight_i = 2^{1-dist_{it}}$ where $dist_{it}$ is the distance to the target term. The run took approximately 7.7 hours on a 16-core system and the compressed model occupies 3.6GB on disk. See Table 4 for an illustration of the similarities captured by the word vectors.

2.5 `word2vec` word vectors

We also applied the `word2vec`[6] implementation of the method proposed by Mikolov et al. (2013) to compute additional vector representations and to induce word clusters. The algorithm is based on neural networks and has been shown to outperform more traditional techniques both in terms of the quality of the resulting representations as well as in terms of computational efficiency. A primary strength of the class of models introduced by Mikolov et al. in comparison to conventional neural network models is that they use a single linear projection layer, thus omitting a number of costly calculations commonly associated with neural networks and making application to much larger data sets than previously proposed methods feasible. We specifically induce 200-dimensional vectors applying the *skip-gram* model with a window size of 5. The model works by predicting the context words within the window focused on each word (see Mikolov et al. for details). Once the vector representation of each word is computed, the words are further clustered with the k-means clustering algorithm with $k = 1000$.

We applied `word2vec` to create three sets of word vectors: one from all PubMed texts, one from all PMC OA texts, and one from the combination of all PubMed and PMC OA texts. For the PubMed and PMC OA subsets, the processing required approx. 12 hours on a 12-core system and

[5] `http://www.nada.kth.se/~xmartin/java/`

[6] `https://code.google.com/p/word2vec/`

Method	Corpus AnEM	BC2GM	NCBID
NERsuite	69.31 / 50.16 / 58.20	74.39 / 75.21 / 74.80	84.41 / 81.69 / 83.02
+ Word clusters	66.43 / 53.11 / 59.03	78.14 / 73.96 / **75.99**	86.91 / 80.12 / **83.38**
Stenetorp et al.	72.90 / 55.89 / **63.27**	74.71 / 66.78 / 70.52	83.86 / 77.84 / 80.73

Table 5: Effect of features derived from `word2vec` word clusters on entity mention tagging (precision/recall/F-score). The best results achieved in a previous evaluation using multiple word representations (Stenetorp et al., 2012) are given for reference.

consumed at peak approx. 4.5GB of memory. The combination of the two took 24 hours and 7.5GB of memory. The resulting vector representations for the three sets are 2-3GB in size. Table 4 shows the nearest words (cosine distance) to selected input words.

3 Extrinsic evaluation

To assess the quality of the word vectors and the clusters created from these vectors, we performed a set of entity mention tagging experiments using three biomedical domain corpora representing various tagging tasks: the BioCreative II Gene Mention task corpus (Smith et al., 2008) (gene and protein names), the Anatomical Entity Mention (AnEM) corpus (Ohta et al., 2012) (anatomical entity mentions) and the NCBI Disease (NCBID) corpus (Doğan and Lu, 2012) (disease names). We compare the results with those of Stenetorp et al. (2012), who previously applied these three corpora in a similar setting to evaluate multiple word representations induced from smaller corpora.

To perform the evaluation, we applied AnatomyTagger (Pyysalo and Ananiadou, 2013), an entity mention tagger using the NERsuite[7] toolkit built on the CRFsuite (Okazaki, 2007) implementation of Conditional Random Fields. For each corpus, we trained one model with default features, and another that augmented the feature set with the cluster ID of each word. We selected hyperparameters ($c2$ and label bias) separately for each corpus and feature set using a grid search with evaluation on the corpus development set. We then trained a final model on the combination of training and development sets, and evaluated it on the test set. We measure performance using exact matching, requiring both tagged mention types and their spans to be precisely correct.[8]

[7] `http://nersuite.nlplab.org`

[8] Note that this criterion is stricter than used in some previous studies on these corpora.

Table 5 shows the extrinsic evaluation results. We find that the word representations are beneficial for tagging performance for all three corpora, improving the performance of a state-of-the-art tagger and surpassing the previously reported results in two out of three cases.

4 Conclusion

We have introduced several resources of general interest to the BioNLP community. First, we assembled a pipeline which fully automatically produces a reference conversion from the complex PubMed and PubMed Central document XML formats into ASCII text suitable for standard text processing tools. Second, we induced 1- to 5-gram models from the entire corpus of over 5 billion tokens. Third, we induced vector space representations using the word2vec and random indexing methods, producing the first word representations induced from the entire available biomedical literature. These can serve as drop-in solutions for BioNLP studies that can benefit from precomputed vector space representations and language models.

In addition to building the resources and making them available, we also illustrated the use of these resources for various named entity recognition tasks. Finally, we have demonstrated the potential of calculating semantic vectors from an existing n-gram based language model using random indexing. All tools and resources introduced in this study are available under open licenses at `http://bio.nlplab.org`.

Acknowledgments

We thank the Chikayama-Tsuruoka lab of the University of Tokyo and the CSC — IT Center for Science of Finland for computational resources and Pontus Stenetorp for input regarding word representations.

References

Nikolas Askitis and Ranjan Sinha. 2007. Hat-trie: a cache-conscious trie-based data structure for strings. In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, pages 97–105.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML 2008*, pages 160–167.

Rezarta Islamaj Doğan and Zhiyong Lu. 2012. An improved corpus of disease mentions in pubmed citations. In *Proceedings of BioNLP 2012*, pages 91–99.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL 2013*.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, Vidas Daudaravicius, and Martin Hassel. 2012. Synonym extraction of medical terms from clinical text using combinations of word space models. In *Proceedings of SMBM 2012*.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL 2012*, pages 873–882.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036. Erlbaum.

Mikael Laakso and Bo-Christer Björk. 2012. Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC medicine*, 10(1):124.

Dekang Lin, Kenneth Ward Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, et al. 2010. New tools for web-scale n-grams. In *LREC*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Andriy Mnih and Geoffrey E Hinton. 2008. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of DSSD 2012*, pages 27–36.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL 2009*, pages 147–155.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Conference of the Cognitive Science Society*, pages 1300–1305.

Larry Smith, Lorraine K Tanabe, Rie J Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2):S2.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL 2012*, pages 1201–1211.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Bionlp Shared Task 2011: Supporting resources. In *Proceedings of BioNLP 2011*, pages 112–120.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi Chikayama. 2012. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. In *Proceedings of SMBM 2012*.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Junichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Advances in informatics*, pages 382–392. Springer.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394.

Karin Verspoor, K Bretonnel Cohen, and Lawrence Hunter. 2009. The textual characteristics of traditional and open access scientific journals are similar. *BMC Bioinformatics*, 10(1):183.

Antonio Jimeno Yepes and Karin Verspoor. 2013. Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material. In *BioLINK SIG 2013*, pages 39–43.

Combining C-value and Keyword Extraction Methods for Biomedical Terms Extraction

Juan Antonio Lossio Ventura, Clement Jonquet
LIRMM, CNRS, Univ. Montpellier 2
Montpellier, France
fName.lName@lirmm.fr

Mathieu Roche, Maguelonne Teisseire
TETIS, Cirad, Irstea, AgroParisTech
Montpellier, France
fName.lName@teledection.fr

Abstract

The objective of this work is to extract and to rank biomedical terms from free text. We present new extraction methods that use linguistic patterns specialized for the biomedical field, and use term extraction measures, such as *C-value*, and keyword extraction measures, such as *Okapi BM25*, and *TFIDF*. We propose several combinations of these measures to improve the extraction and ranking process. Our experiments show that an appropriate harmonic mean of *C-value* used with keyword extraction measures offers better precision results than used alone, either for the extraction of single-word and multi-words terms. We illustrate our results on the extraction of English and French biomedical terms from a corpus of laboratory tests. The results are validated by using UMLS (in English) and only MeSH (in French) as reference dictionary.

1 Introduction

Language evolves faster than our ability to formalize and catalog concepts or possible alternative terms of these concepts. This is even more true for French in which the number of terms formalized in terminologies is significantly less important than in English. That is why our motivation is to improve the precision of automatic terms extraction process. Automatic Term Recognition (ATR) is a field in language technology that involves the extraction of technical terms from domain-specific language corpora (Zhang et al., 2008). Similarly, Automatic Keyword Extraction (AKE) is the process of extracting the most relevant words or phrases in a document with the propose of automatic indexing. Keywords, which we define as a sequence of one or more words, provide a compact representation of a document's content; two popular AKE measures are *Okapi BM25* (Robertson et al., 1999) and *TFIDF* (also called weighting measures). These two fields are summarized in Table 1.

	ATR	**AKE**
Input	one large corpus	single document
Output	terms of a domain	keywords of a doc
Domain	very specific	none
Exemples	*C-value*	*TFIDF*, *Okapi*

Table 1: Differences between ATR and AKE.

In our work, we adopt as baselines an ATR method, *C-value* (Frantzi et al., 2000), and the best two AKE methods (Hussey et al., 2012), previously mentioned and considered state-of-the-art. Indeed, the C-value, compared to other ATR methods, often gets best precision results and specially in biomedical studies (Knoth et al., 2009), (Zhang et al., 2008), (Zhang et al., 2004). Moreover, *C-value* is defined for multi-word term extraction but can be easily adapted for single-word term and it has never been applied to French biomedical text, which is appealing in our case.

Our experiments present a great improvement of the precision with these new combined methods. We give priority to precision in order to focus on extraction of new valid terms (i.e., for a candidate term to be a valid biomedical term or not) rather than on missed terms (recall).

The rest of the paper is organized as follows: Section 2 describes the related work in the field of ATR, and specially the uses of the *C-value*; Section 3 presents our combination of measures for ranking candidate terms; Section 4 shows and discusses our experiment results; and Section 5 concludes the paper.

2 Related work

ATR studies can be divided into four main categories: (i) rule-based approaches, (ii) dictionary-based approaches, (iii) statistical approaches, and (iv) hybrid approaches. Rule-based approaches for

instance (Gaizauskas et al., 2000), attempt to recover terms thanks to the formation patterns, the main idea is to build rules in order to describe naming structures for different classes using orthographic, lexical, or morphosyntactic characteristics. Dictionary-based approaches use existing terminology resources in order to locate term occurrences in texts (Krauthammer et al., 2004). Statistical approaches are often built for extracting general terms (Eck et al., 2010); the most basic measure is *frequency*. *C/NC-value* (Frantzi et al., 2000), is another statistical method well known in the literature that combines statistical and linguistic information for the extraction of multi-word and nested terms. While most studies address specific types of entities, *C/NC-value* is a domain-independent method. It was also used for recognizing terms from biomedical literature (Hliaoutakis et al., 2009). The *C/NC-value* method was also applied to many different languages besides English (Frantzi et al., 2000) such as Japanese (Mima et al., 2001), Serbian (Nenadić et al., 2003), Slovenian (Vintar, 2004), Polish (Kupsc, 2006), Chinese (Ji et al., 2007), Spanish (Barrón et al., 2009), and Arabic (Khatib et al., 2010), however to the best of our knowledge not to French. An objective of this work is to combine this method with AKE methods and to apply the combined measures to English and French. We believe that the combination of biomedical term extraction and the extraction of keywords describing a document, could be beneficial since keywords techniques give greater importance to the actual terms of this domain. This combination has never been proposed and experimented in the literature.

3 Proposed Methodology for Automatic Biomedical Term Extraction

This section describes the baselines measures and their customizations as well as the new combinations of these measures that we propose for automatic biomedical terms extraction and ranking. Our method for automatic term extraction has four main steps: (1) Part-of-Speech tagging, (2) Candidate terms extraction,(3) Ranking of candidate terms, (4) Computing of new combined measures.

Note, *C-value* is a method that deals with an unique corpus as input whereas AKE methods deal with several documents (cf. Table 1) then we need to do the union of documents for *C-value* to consider the whole corpus as an unique document. A preliminary step is the creation of patterns for French and English, as described hereafter.

3.1 Part-of-Speech tagging

Part-of-speech (POS) tagging is the process of assigning each word in a text to its grammatical category (e.g., noun, adjective). This process is performed based on the definition of the word or on the context which it appears in.

We apply part-of-speech to the whole corpus. We evaluated three tools (TreeTagger, Stanford Tagger and Brill's rules), and finally choose TreeTagger which gave best results and is usable both for French and English.

3.2 Candidate terms extraction

As previously cited work, we supposed that biomedical terms have similar syntactic structure. Therefore, we build a list of the most common lexical patterns according the syntactic structure of biomedical terms present in the UMLS[1] (for English) and the French version of MeSH[2] (for French). We also do a part-of-speech tagging of the biomedical terms using TreeTagger[3], then compute the frequency of syntactic structures. We finally choose the 200 highest frequencies to build the list of patterns for each language. The number of terms used to build these lists of patterns was 2 300 000 for English and 65 000 for French.

Before applying measures we filter out the content of our input corpus using patterns previously computed. We select only the candidate terms which syntactic structure is in the patterns list.

3.3 Ranking of candidate terms

3.3.1 Using *C-value*

The *C-value* method combines linguistic and statistical information (Frantzi et al., 2000); the linguistic information is the use of a general regular expression as linguistic patterns, and the statistical information is the value assigned with the *C-value* measure based on frequency of terms to compute the *termhood* (i.e., the association strength of a term to domain concepts). The aim of the *C-value* method is to improve the extraction of nested terms, it was specially built for extracting multi-word terms.

$$C\text{-}value(a) = \begin{cases} w(a) \times f(a) & \text{if } a \notin nested \\ w(a) \times \left(f(a) - \frac{1}{|S_a|} \times \sum_{b \in S_a} f(b) \right) & \\ \text{otherwise} & \end{cases} \quad (1)$$

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

[2] http://mesh.inserm.fr/mesh/

[3] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger

Where a is the candidate term, $w(a) = \log_2(|a|)$, $|a|$ the number of words in a, $f(a)$ the frequency of a in the unique document, S_a the set of terms that contain a and $|S_a|$ the number of terms in S_a. In a nutshell, *C-value* either uses frequency of the term if the term is not include in other terms (first line), or decrease this frequency if the term appears in other terms, by using the frequency of those other terms (second line).

We modified the measure in order to extract all terms (single-word + multi-words terms), as suggested in (Barrón et al., 2009) in different manners: in the formula $w(a) = \log_2(|a|)$, we use $w(a) = \log_2(|a| + 1)$ in order to avoid null values (for single-word terms). Note that we do not use a stop word list nor a threshold for frequency as it was originally proposed.

3.3.2 Using *Okapi - TFIDF*

Those measures are used to associate each term of a document with a weight that represents its relevance to the meaning of the document it appears relatively to the corpus it is included in. The output is a ranked list of terms for each document, which is often used in information retrieval, to order documents by their importance given a query (Robertson et al., 1999). *Okapi* can be seen as an improvement of *TFIDF* measure, taking into account the document length.

The outputs of *Okapi* and *TFIDF* are calculated with a variable number of data so their values are heterogeneous. To manipulate these lists, the weights obtained from each document must be normalized. Once values normalized we have to merge the terms into a single list unique for the whole corpus to compare the results. Clearly the precision will depend on the method used to perform such merging. We merged following three functions, which calculate respectively the sum(S), max(M) and average(A) of the measures values of the term in whole the corpus. At the end of this task we have three lists from *Okapi* and three lists from *TFIDF*. The notation for these lists are $Okapi_X(a)$ and $TFIDF_X(a)$, where a is the term, X the factor $\in \{M, S, A\}$. For example, $Okapi_M(a)$ is the value obtained by taking the maximum Okapi value for a term a in the whole corpus.

3.4 Computing the New Combined Measures

With the goal of improving the precision of terms extraction we have conceived two new combined measures schemes, described hereafter, taking into account the values obtained in the above steps.

3.4.1 F-OCapi and F-TFIDF-C

Considered as the harmonic mean of the two used values, this method has the advantage of using all the values of the distribution.

$$F\text{-}OCapi_X(a) = 2 \times \frac{Okapi_X(a) \times C\text{-}value(a)}{Okapi_X(a) + C\text{-}value(a)} \quad (2)$$

$$F\text{-}TFIDF\text{-}C_X(a) = 2 \times \frac{TFIDF_X(a) \times C\text{-}value(a)}{TFIDF_X(a) + C\text{-}value(a)} \quad (3)$$

3.4.2 C-Okapi and C-TFIDF

Our assumption is that *C-value* can be more representative if the frequency, in Equation (1), of the terms is replaced with a more significant value, in this case the *Okapi's* or *TFIDF's* values of the terms (over the whole corpus).

$$C\text{-}m_X(a) = \begin{cases} w(a) \times m_X(a) & \text{if } a \notin nested \\ w(a) \times \left(m_X(a) - \frac{1}{|S_a|} \times \sum_{b \in S_a} m_X(b) \right) & \text{otherwise} \end{cases}$$

Where $m_X(a) = \{Okapi_X | TFIDF_X\}$, and $X \in \{M, S, A\}$.

4 Experiments and Results

4.1 Data and Experimental Protocol

We used biological laboratory tests, Labtestonline.org, as *corpus*. This site provides information in several languages to patient or family caregiver on clinical lab tests. Each test which forms a document in our corpus, includes the *formal lab test name*, some *synonyms* and possible *alternate names* as well as a description of the test. Our extracted corpus contains 235 clinical tests (about 400 000 words) for English and 137 (about 210 000 words) for French.

To automatically validate our candidate terms we compute a validation dictionary that include the *official name*, the *synonyms* and *alternate names* of the labtestonline tests plus all UMLS terms for English and the MeSH terms for French. These terminologies are references in the domain therefore each extracted term found in those is validated as a true term. Note that as a consequence we obtain 100% *Recall* with the whole list of extracted terms.

4.2 Experiments and results

Results are evaluated in terms of *Precision* obtained over the top k terms at different steps of our work presented in previous section. $Okapi$ and $TFIDF$ provided three lists of ranked candidate terms (M, S, A). For each combined measure using $Okapi$ or $TFIDF$, the experiments are done with the three lists. Therefore, the number of ranked list to compare is $C\text{-}value(1) + Okapi(3) + TFIDF(3) + F\text{-}OCapi(3) + F\text{-}TFIDF\text{-}C(3) + C\text{-}Okapi(3) + C\text{-}TFIDF(3) = 19$. In addition we experimented either for all (single and multi) or multi terms which finally give 38 ranked lists. Then, we select all terms (single and multi) or only muli-terms ($19 \times 2 = 38$ experiments for each language).

The following lines show part of the experiment results done all or multi terms, only and considering the top 60, 300 and 900 extracted terms, because it is appropriate and easier for an expert to evaluate the first best extracted terms. Table 2 and Table 3 compare the precision between the best baselines measures and the best combined measures. Best results were obtained in general with $F\text{-}TFIDF\text{-}C_M$ for English and $F\text{-}OCapi_M$ for French. **These tables prove that the combined measures based on the harmonic mean are better than the baselines measures, and specially for multi word terms, for which the gain in precision reaches 16%**. This result is particularly positive because in the biomedical domain it is often more interesting to extract multi-word terms than single-word terms. However, one can notice that results obtained to extract all terms with $C\text{-}Okapi_S$ and $C\text{-}TFIDF_S$ are not better than $Okapi_X$ or $TFIDF_X$ used directly. The reason is because the performance of those new combined measures is affected when single word terms are extracted. Definitively, the new combined measures are really performing for multi word term.

Results of AKE methods for English show that $TFIDF_X$ obtains better results than $Okapi_X$. The main reason for this, is because the size of the English corpus is larger than the French one, and *Okapi* is known to perform better when the corpus size is smaller (Lv et al., 2011).

In addition, Table 3 shows that *C-value* can be used to extract French biomedical terms with a better precision than what has been obtained in previous cited works with different languages. The precision of *C-value* for the previous work was between 26% and 31%.

	All Terms			Multi Terms		
	60	300	900	60	300	900
$Okapi_M$	0.96	0.95	0.82	0.68	0.62	0.54
$Okapi_S$	0.83	0.89	0.85	0.58	0.57	0.55
$Okapi_A$	0.72	0.31	0.27	0.48	0.39	0.26
$TFIDF_M$	0.97	0.96	0.84	0.71	0.63	0.54
$TFIDF_S$	0.96	0.95	**0.93**	0.82	0.71	0.61
$TFIDF_A$	0.78	0.74	0.63	0.50	0.40	0.37
$C\text{-}value$	0.88	0.92	0.89	0.72	0.71	0.62
$F\text{-}OCapi_M$	0.73	0.87	0.84	0.79	0.69	0.58
$F\text{-}TFIDF\text{-}C_M$	**0.98**	**0.97**	0.86	**0.98**	**0.73**	**0.65**
$C\text{-}Okapi_S$	0.88	0.86	0.80	0.61	0.58	0.53
$C\text{-}TFIDF_S$	0.96	0.95	0.86	0.85	0.71	0.61

Table 2: Extract of precision comparison for term extraction for English.

	All Terms			Multi Terms		
	60	300	900	60	300	900
$Okapi_M$	**0.90**	0.61	0.37	0.53	0.31	0.18
$Okapi_S$	0.30	0.31	0.37	0.23	0.30	0.37
$Okapi_A$	0.52	0.31	0.16	0.30	0.17	0.16
$TFIDF_M$	0.75	0.51	0.37	0.45	0.28	0.18
$TFIDF_S$	0.68	0.48	0.42	0.53	0.33	0.22
$TFIDF_A$	0.12	0.39	0.29	0.17	0.16	0.11
$C\text{-}value$	0.43	0.42	0.43	0.35	0.34	**0.26**
$F\text{-}OCapi_M$	0.73	**0.62**	**0.43**	**0.65**	**0.35**	0.22
$F\text{-}TFIDF\text{-}C_M$	0.85	0.57	0.39	0.62	0.31	0.19
$C\text{-}Okapi_S$	0.28	0.32	0.34	0.23	0.28	0.20
$C\text{-}TFIDF_S$	0.65	0.55	0.38	0.50	0.32	0.19

Table 3: Extract of precision comparison for term extraction for French.

We also have done experiments with two more corpus: (i) the Drugs data from MedlinePlus[4] in English and, (ii) PubMed[5] citations' titles in English and French, we have verified that the new combined measures are performing better, particularly these based on the harmonic mean, $F\text{-}TFIDF\text{-}C_M$ and $F\text{-}OCapi_M$.

5 Conclusions and Perspectives

This work present a methodology for term extraction and ranking for two languages, French and English. We have adapted *C-value* to extract French biomedical terms, which was not proposed in the literature before. We presented and evaluated two new measures thanks to the combination of three existing methods. The best results were obtained by combining *C-value* with the best results from AKE methods, i.e., $F\text{-}TFIDF\text{-}C_M$ for English and $F\text{-}OCapi_M$ for French.

For our future evaluations, we will enrich our dictionaries with BioPortal's[6] terms for English and CISMeF's[7] terms for French. Our next task will be the extraction of relations between these new terms and already known terms, to help in ontology population. In addition, we are currently implementing a web application that implements these measures for the community.

[4] http://www.nlm.nih.gov/medlineplus/

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

[6] http://bioportal.bioontology.org/

[7] http://www.chu-rouen.fr/cismef/

Acknowledgments

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University Montpellier 2 and CNRS.

References

Alberto Barrón-Cedeño, Gerardo Sierra, Patrick Drouin, Sophia Ananiadou. 2009. An Improved Automatic Term Recognition Method for Spanish. *Proceeding of Computational Linguistics and Intelligent Text Processing,* pp 125-136.

NeesJan Eck, Ludo Waltman, EdC.M. Noyons, ReindertK Buter. 2010. Automatic term identification for bibliometric mapping. *SpringerLink, Scientometrics,* Volume 82, Number 3.

Katerina Frantzi, Sophia Ananiadou, Hideki Mima 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal of Digital Libraries, 3(2)* pp.117-132.

Robert Gaizauskas, George Demetriou, Kevin Humphreys. 2000. Term Recognition and Classification in Biological Science Journal Articles. *Proceedings of the Computational Terminology for Medical and Biological Applications Workshop*, pp 37–44.

Angelos Hliaoutakis, Kaliope Zervanou and Euripides G.M. Petrakis. 2009. The AMTEx approach in the medical document indexing and retrieval application. *Data and Knowledge Eng.,* pp 380-392.

Richard Hussey, Shirley Williams, Richard Mitchell. 2012. Automatic keyphrase extraction: a comparison of methods. *Proceedings of the International Conference on Information Process, and Knowledge Management,* pp. 18-23.

Luning Ji, Mantai Sum, Qin Lu, Wenjie Li, Yirong Chen. 2007. Chinese Terminology Extraction Using Window-Based Contextual Information. *Proceeding of CICLing, LNCS,* pp.62–74.

Khalid Al Khatib, Amer Badarneh. 2010. Automatic extraction of Arabic multi-word terms. *Proceeding of Computer Science and Information Technology.* pp 411-418.

Petr Knoth, Marek Schmidt, Pavel Smrz, Zdenek Zdrahal. 2009. Towards a Framework for Comparing Automatic Term Recognition Methods. *Conference Znalosti*.

Michael Krauthammer, Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics,* pp 512–526.

Anna Kupsc. 2006. Extraction automatique de termes à partir de textes polonais. *Journal Linguistique de Corpus.*

Yuanhua Lv, ChengXiang Zhai. 2011. When documents are very long, BM25 fails! *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval,* pp.1103–1104.

Olena Medelyan, Eibe Frank, Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. *Proceeding of the International Conference of Empirical Methods in Natural Language Processing, Singapore*.

Hideki Mima, Sophia Ananiadou, . 2001. An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese. *Japanese Term Extraction. Special issue of Terminology,* vol 6:2

Goran Nenadić, Irena Spasić, Sophia Ananiadou. 2003. Morpho-syntactic clues for terminological processing in Serbian. *Proceeding of the EACL Workshop on Morphological Processing of Slavic Languages,* pp.79–86.

Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas B. Griffith, Clement Jonquet, Daniel L. Rubin, Margaret-Anne Storey, Christopher G. Chute, Mark A. Musen. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research,* pp. 170-173 vol. 37.

Stephen Robertson, Steve Walker, Micheline Hancock-Beaulieu. 1999. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. *IN*. pp. 253–264 vol. 21.

Francesco Sclano, Paola Velardi. 2007. TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. *In Enterprise Interoperability II,* pp. 287-290.

Špela Vintar. 2004. Comparative Evaluation of C-Value in the Treatment of Nested Terms. *Workshop (Methodologies and Evaluation of Multiword Units in Real-world Applications),* pp.54–57.

Yongzheng Zhang, Evangelos Milios, Nur Zincirheywood. 2004. A Comparison of Keyword- and Keyterm-Based Methods for Automatic Web Site Summarization. *AAAI04 Workshop on Adaptive Text Extraction and Mining,* pp. 15–20.

Ziqi Zhang, José Iria, Christopher Brewster, Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Open Information Extraction from Biomedical Literature Using Predicate-Argument Structure Patterns

Nhung T. H. Nguyen[†*], Makoto Miwa[‡], Yoshimasa Tsuruoka[†] and Satoshi Tojo[*]
[†]The University of Tokyo, 3-7-1 Hongo, Bunkyo-ku, Tokyo, Japan
{nhung, tsuruoka}@logos.t.u-tokyo.ac.jp
[‡]The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK
makoto.miwa@manchester.ac.uk
[*]Japan Advanced Institue of Science and Technology, Ishikawa, Japan
{nthnhung, tojo@jaist.ac.jp}

Abstract

In this paper, we propose an open information extraction (Open IE) system, which attempts to extract relations (or facts) of any type from biomedical literature. What distinguishes our system from existing Open IE systems is that it uses predicate-argument structure patterns to detect the candidates of possible biomedical facts. We have manually evaluated the output of our system and found that it is reasonably accurate (50% precision). We have also applied our system to the whole MEDLINE and revealed that the relations between 'Amino Acid, Peptide, or Protein' entities are the most frequently described type of relations.

1 Introduction

Relation extraction is one of the most important tasks in biomedical text mining. Most of the studies on this topic have focused on specific or predefined types of relations, such as protein-protein interaction (Yakushiji et al., 2006; Airola et al., 2008; Miwa et al., 2009), drug-drug interaction (Segura-Bedmar et al., 2013), and biomolecular events (Nédellec et al., 2013). The scope of the types of relations that can be extracted by existing approaches is, therefore, inherently limited.

Recently, an information extraction paradigm called Open Information Extraction (Open IE) has been introduced to overcome the above-mentioned limitation (Banko et al., 2007; Fader et al., 2011; Mausam et al., 2012). Open IE systems aim to extract all possible relations from text. Although the concept of Open IE is certainly appealing, we have found that state-of-the-art Open IE systems, namely Reverb (Fader et al., 2011) and OLLIE (Mausam et al., 2012), do not perform well on biomedical text – they can capture relational phrases with reasonable accuracy but often fails to correctly identify their arguments.

This observation has motivated us to develop an Open IE system specifically designed for biomedical texts. Our system uses Predicate-Argument Structures (PAS) patterns to detect the candidates of possible biomedical facts. We decided to use PAS patterns because they are well normalized forms that represent deep syntactic relations. In other words, multiple syntactic variations are reduced to a single PAS, thereby allowing us to cover many kinds of expressions with a small number of PAS patterns. We first apply an HPSG-based syntactic parser to input sentences, and then match its output to predefined PAS patterns to detect pairs of relevant noun phrases (NPs). Named entities in these pairs are then detected; and finally, relations between these entities are extracted. The output of our system is, hopefully, a set of all semantic relations contained in the input.

Our contribution in this paper is twofold: (1) a simple but effective set of syntactic patterns for general relation extraction, and (2) an Open IE system that extracts biomedical facts from biomedical text; to the best of our knowledge, our system is the first Open IE system that attempts to detect relations from the whole MEDLINE in a general schema.

2 Related Work

Banko et al. (2007) introduced Open IE as a novel paradigm that facilitates domain independent discovery of relations extracted from text and readily scales to the diversity and size of the Web corpus. The system detects the tuples in the format of (argument 1; relational phrase; argument 2) without a pre-specified set of relations or domain-specific knowledge engineering. Several Open IE systems have been proposed up to now, including TextRunner (Banko et al., 2007), ReVerb (Fader et al., 2011), OLLIE (Mausam et al., 2012).

In the biomedical domain, large-scale event extraction has attracted many researchers (Rindflesch and Fiszman, 2003; Miyao et al., 2006; Björne et al., 2010; Taura et al., 2010; Rindflesch et al., 2011; Kilicoglu et al., 2012; Van Landeghem et al., 2013). Miyao et al. (2006) propose a system that extracts verb-mediated relations between genes, gene products, and diseases from MEDLINE. The output of their system is served as a database for MEDIE (Ohta et al., 2010), a semantic search engine on MEDLINE. Björne et al. (2010) apply their system to the titles and abstracts of all PubMed citations. Kilicoglu et al. (2012) also run their system on the entire set of PubMed citations to create SemMedDB, a repository of semantic predications.

SemRep (Rindflesch and Fiszman, 2003; Rindflesch et al., 2011) extracts semantic relationships from the titles and abstracts of all PubMed citations. Their relationships are represented by 30 specific *predicates* restricted to a limited number of verbs. Nebot and Berlanga (2012) extracted explicit binary relations of the form <*subject, predicate, object*> from CALBC initiative. To detect candidate relations, they proposed seven simple lexico-syntactic patterns. These two systems perform general relations extraction similar to ours, but unlike our system, neither of them use PAS patterns.

3 Our Open IE Framework

Since we focus on a general schema of relations, there is no labeled corpus suitable for learning the extraction model. Our system, therefore, relies solely on the input text and its linguistic characteristics such as the form, meaning, and context of the words. More specifically, we create patterns to capture these characteristics of text and then extract relations.

In order to find appropriate PAS patterns, we have first observed textual expressions that represent biomedical relations in GENIA corpus and found that those relations are usually expressed with verbs and prepositions; for example, $Entity_A$ {*affect, cause, express, inhibit ...*} $Entity_B$ or $Entity_A$ {*arise, happen, ...*} {*in, at, on ...*} *Location*. Our patterns in predicate-argument form and their corresponding examples are presented in Table 1. Patterns 1, 2, 3 and 4 are presented for transitive verbs. Intransitive verbs are captured by Pattern 5. The final pattern (Pattern 6) is used for prepositions, which would capture localization and whole-part relations. The elements NP_1 and NP_2 in each pattern are considered as candidate relations. In our system, Enju, an HPSG parser (Matsuzaki et al., 2007; Miyao et al., 2008), is employed to extract these candidates.

We then apply MetaMap[1] (Aronson and Lang, 2010) to identify named entities in the extracted NP pairs. At this stage, we apply two post-processes to remove false positive output from MetaMap. In the first process, we remove all entities that are verbs, adjectives, prepositions or numbers because we are only interested in noun or noun phrase ones. The second post-process is used to avoid common noun entities, such as 'binding', 'behaviors' and 'kinds'. In this process, we apply MetaMap to the whole MEDLINE and construct a dictionary of named entities and their occurrences. We then remove highly frequent entities from the dictionary. This dictionary is used to check the validity of named entities. Our statistical results on the whole MEDLINE revealed that the post-processes filtered out **70.83%** of the entities extracted by MetaMap. This filtering will help our system avoid extracting irrelevant relations.

After the above two post-processes, we obtain named entities in relevant NP pairs. Let us denote by $<NP_1, NP_2>$ a relevant NP pair, by e_{1i} ($i = 1, 2, ...$) entities in NP_1, and by e_{2j} ($j = 1, 2, ...$) entities in NP_2. Since NP_1 and NP_2 are relevant, we assume that every pair of entities $<e_{1i}, e_{2j}>$ is relevant, which means that they constitute a semantic relation. However, this assumption is so strong that it may create incorrect relations. In order to improve the precision of our system, we use the UMLS semantic network[2] as a constraint in extracting semantic relations. Let us denote by $<s_1, s_2>$ the pair of semantic types of $<e_{1i}, e_{2j}>$. If and only if $<s_1, s_2>$ exists in this semantic network, $<e_{1i}, e_{2j}>$ can constitute a relation.

4 Experimental Results

4.1 Performance on General Relations

Since there is no available labeled corpus for a general schema of relations, we manually evaluated our system on our own test set. This test set was created by randomly selecting 500 sentences from MEDLINE. Our system was given this set

[1] We employed MetaMap 2012 version 2 from `http://metamap.nlm.nih.gov/#Downloads`

[2] `http://semanticnetwork.nlm.nih.gov/`

No.	Type	PAS Patterns	Examples
1	Verb	$NP_1 \leftarrow \mathbf{Verb} \rightarrow NP_2$	protein RepA(cop) ← affects →a single amino acid
2		$NP_1 \leftarrow \mathbf{Verb} \rightarrow by + NP_2$	Diabetes ← induced → by streptozotocin injection
3		$NP_1 \leftarrow \mathbf{Verb} \rightarrow NP'$ $\uparrow$ $Prep. \rightarrow NP_2$	Endothelin-1 (ET-1) and ET-3 ← had → a strong effect $\uparrow$ $in \rightarrow$ all trabeculae
4		$NP_1 \leftarrow \mathbf{be} \rightarrow ADJP \leftarrow Prep. \rightarrow NP_2$	EPO receptor ← be → present ← in → tubular epithelial cells
5		$NP_1 \leftarrow \mathbf{Verb} \leftarrow Prep. \rightarrow NP_2$	subacute hepatitis ← results ← from → intravenous drug use
6	Prep.	$NP_1 \leftarrow \mathbf{Prep}. \rightarrow NP_2$	vitronectin ← in → the connective tissue

Table 1: Our PAS patterns focus on verb and preposition predicates. An arrow goes from a to b means a modifies b and a is called a predicate, b is called an argument. $<NP_1, NP_2>$ is a relevant NP pair in each pattern.

	Conf.	# of Rel.	Precision
ReVerb	≥0.3	75	46.67
	≥0.5	72	47.22
	≥0.7	58	46.55
OLLIE	≥0.3	124	38.71
	≥0.5	114	41.22
	≥0.7	89	42.69
Our patterns	-	**438**	**50.00**

Table 2: The precisions of relation extraction on our test set when using ReVerb and OLLIE with three confidence scores of 0.3, 0.5 and 0.7, and our PAS patterns to extract NP pairs.

as input, and returned a set of binary relations as output.

For comparison, we conducted experiments using two state-of-the-art Open IE systems, namely, ReVerb (Fader et al., 2011) and OLLIE (Mausam et al., 2012). We employed these two systems to extract relevant NP pairs in place of our PAS patterns. We chose confidence scores of 0.3, 0.5 and 0.7 as the thresholds for accepting generated tuples as candidate relations in our experiments. Next, the other processes were applied in the same way as our system. We report our evaluation results in Table 2. Compared with ReVerb and OLLIE, our PAS patterns generated the highest number of relations with the highest precision. This indicates that our PAS patterns perform better than the other approaches.

The causes of false positive relations include MetaMap errors, parser errors, and our greedy extraction. Since our system is based on the Enju parser, if the parser captures wrong noun phrases, our system will generate incorrect relevant pairs. For example, with this input "{[Laminin]}$_{NP_1}$ was *located in* {the zone of the basal [membrane], whereas [tenascin] was mainly found in the mucosal [vessels]}$_{NP_2}$", based on the NP pair $<NP_1, NP_2>$, the system returned two relations r_1 (Laminin, membrane) and r_2 (Laminin, vessels). In this example, the parser failed to detect the second NP of the pair; the correct one should be 'the zone of the basal membrane', not including 'whereas' clause. This error caused a false positive relation of *(Laminin, vessels)*. Extracted relation r_1 (Laminin, membrane) is also not correct because of the MetaMap error, i.e., the entity 'membrane' should be 'basal membrane'.

Although we use the Semantic Network to limit the generated relations, there are several false positive ones. For instance, given an input sentence: "{Efficiency of presentation of a peptide epitope by a [MHC class I molecule]}$_{NP_1}$ *depends on* {two parameters: its binding to the [MHC] molecule and its generation by intracellular Ag processing}$_{NP_2}$", the pair $<NP_1, NP_2>$ created a relation of (MHC class I molecule, MHC). This relation resulted from our greedy extraction. However, it is incorrect because 'MHC class I molecule' or 'MHC' is not the main subject of this sentence.

Table 2 shows that when using ReVerb and OLLIE to generate NP pairs, the numbers of extracted relations are significantly lower than those when using our patterns. The main reason is that these systems have failed to capture NP pairs in many sentences. In our test set, ReVerb and OLLIE could not extract NP pairs from 150 sentences and 95 sentences respectively; while our system could not extract pairs from 14 sentences only. Given the input sentence "{[Total protein], [lactate dehydrogenase] (LDH), [xanthine oxidase] (XO), [tumor necrosis factor] (TNF), and [interleukin 1] (IL-1)}$_{NP_1}$ were *measured in* {[bronchoalveolar lavage fluid] (BALF)}$_{NP_2}$.", ReVerb and OLLIE cannot extract any tuples, while our system generated a NP pair of $<NP_1, NP_2>$ and returned five

	AIMed		BioInfer		LLL	
	Pre.	Re.	Pre.	Re.	Pre.	Re.
(1)	71.8	*48.4*	-	-	-	-
(2)	52.9	61.8	47.7	59.9	72.5	87.2
(3)	55.0	68.8	65.7	71.1	77.6	86.0
Our system	30.3	**52.5**	51.2	44.9	87.5	81.5

Table 3: Performance of our system on AIMed, BioInfer and LLL corpora, compared with some notable systems for PPI: (1) Yakushiji et al. (2006), (2) Airola et al. (2008), and (3) Miwa et al. (2009).

	MedLine		DrugBank	
	Pre.	Re.	Pre.	Re.
Best system	55.8	*50.5*	81.6	83.8
Worst system	*62.5*	42.1	38.7	73.9
Our PAS patterns	27.0	**62.5**	41.0	61.6

Table 4: Performance of our system on MedLine and DrugBank corpora of SemEval-2013 Task 9 (Segura-Bedmar et al., 2013), compared with the best and worst system in that shared task.

correct relations between 'bronchoalveolar lavage fluid' and five entities in NP_1. This is a representative example showing the advantage of our PAS patterns in extracting candidate relations.

4.2 Performance on Predefined Relations

We also conducted experiments to check if our PAS patterns could cover other predefined relations, including Protein-Protein Interaction (PPI) and Drug-Drug Interaction (DDI). Regarding PPI, we applied our patterns to AIMed, BioInfer and LLL (Airola et al., 2008; Pyysalo et al., 2008). The available gold standard entities in these corpora were used instead of MetaMap output. Our experimental results and the results of some machine learning-based systems on PPI are shown in Table 3. It should be noted that these systems were evaluated by using 10-fold cross validation or using the test set; while our method is rule-based and thus we simply applied our patterns to the whole labeled corpora.

We conducted the same experiment for DDI on the SemEval-2013 task 9 corpus (Segura-Bedmar et al., 2013) and report the results in Table 4.

Results in Table 3 and Table 4 show that although our PAS patterns are very simple, their performance is competitive with other machine learning methods on both PPI and DDI. In some cases, our method even outperforms the other ones such as PPI on AIMed corpus and DDI on MedLine corpus in recall.

Rank	Semantic Relation		Count
	Entity 1	Entity 2	
1	aapp	aapp	**2,006,301**
2	cell	aapp	1,770,561
3	bpoc	aapp	1,046,523
4	gngm	aapp	1,008,017
5	dsyn	dsyn	909,195
6	aapp	dsyn	869,143
7	aapp	bacs	680,349
8	bpoc	mamm	676,325
9	lbpr	aapp	650,571
10	bpoc	dsyn	626,644

Table 5: The ten most frequent types of semantic relations found in the whole MEDLINE.

4.3 Extracting Semantic Relations in MEDLINE

We have applied our system to the whole MEDLINE[3] to extract semantic relations and calculated their frequencies to see which relations are common in this corpus. The statistical results in Table 5 show that the most common semantic relation in MEDLINE is the relation between 'Amino Acid, Peptide or Protein' (aapp) entities[4]. This explains why researchers in BioNLP have been focusing on protein-protein interaction. We can also see that 'Amino Acid, Peptide or Protein' entities contribute in 7 over 10 most popular relations, which shows their important role in the biomedical domain.

5 Conclusion

In this work, we have developed an Open IE system for biomedical literature by employing six PAS patterns to extract the candidates of possible biomedical facts. The system extracted 438 relations from our test set and 50% of those were correct. Compared with ReVerb and OLLIE, our patterns have presented better performance in extracting relevant NP pairs. The experimental results show that our patterns are effective on both general and specific relations. The statistical analysis on the result of the whole MEDLINE provides support for the intuition that the most common semantic relations are the ones between 'Amino Acid, Peptide and Protein' entities.

[3]The version used in this paper is the 2012 MEDLINE/PubMed baseline database.

[4]The semantic types of entities in Table 5 are in short form for our convenience, for their full form, please refer to `http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF`

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 1–9.

Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 17(3):229–236.

Michele Banko, Michael Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI*, pages 2670–2676.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Scaling up Biomedical Event Extraction to the Entire Pubmed. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP'10)*, pages 28–36. ACL.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP*, pages 1535–1545. ACL.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. 2012. SemMedDB: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2007. Efficient HPSG parsing with supertagging and cfg-filtering. In *Proceedings of IJCAI*, pages 1671–1676.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *Proceedings of EMNLP-CoNLL*, pages 523–534. ACL.

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2009. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *I. J. Medical Informatics*, 78(12):39–46.

Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of ACL*.

Yusuke Miyao, Rune Sætre, Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of ACL*, pages 46–54.

Victoria Nebot and Rafael Berlanga. 2012. Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. *Knowledge and Information Systems*.

C. Nédellec, R. Bossy, J.-D. Kim, J-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum. 2013. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, August.

Tomoko Ohta, Takuya Matsuzaki, Naoaki Okazaki, Makoto Miwa, Rune Stre, Sampo Pyysalo, and Jun'ichi Tsujii. 2010. Medie and info-pubmed: 2010 update. *BMC Bioinformatics*, 11(S-5):P7.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(S-3).

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

Thomas C. Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, and Dongwook Shin. 2011. Semantic MEDLINE: An advanced information management application for biomedicine. *Information Services & Use*, (31):15–21.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. SemEval-2013 task 9 : Extraction of Drug-Drug interactions from Biomedical Texts. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, June.

Kenjiro Taura, Takuya Matsuzaki, Makoto Miwa, Yoshikazu Kamoshida, Daisaku Yokoyama, Nan Dun, Takeshi Shibata, Choi Sung Jun, and Jun'ichi Tsujii. 2010. Design and implementation of GXP make - a workflow system based on make. In *eScience*, pages 214–221. IEEE Computer Society.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, 8(4).

Akane Yakushiji, Yusuke Miyao, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2006. Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *Proceedings of EMNLP*, pages 284–292. ACL.

Sharing Reference Texts for Interoperability of Literature Annotation

Jin-Dong Kim
Database Center for Life Science (DBCLS)
Research Organization of Information and Systems (ROIS)
2-11-16, Yayoi, Bunkyo-ku, Tokyo, Japan
jdkim@dbcls.rois.ac.jp

Abstract

As an effort to improve the interoperability of literature annotation, the paper suggests to share the reference texts whereon annotations produced by different projects may be aligned. PubAnnotation is a web-based public system implemented to support the idea. Its two key features, (1) reference text provision and (2) annotation alignment, are presented in the paper.

1 Introduction

Corpus annotation is considered indispensable for the development of text mining technology. In the case of the area of life sciences, due to the rapid increasing rate of publications, the need for text mining is very high. In the area, thanks to the existence of the extensive databases of literature, e.g., PubMed[1] and PubMed Central (PMC)[2], annotation efforts have been largely made to the publicly accessible portion of literature, e.g., PubMed abstracts and the Open Access subset of PMC (OAPMC)[3].

Given them, in the area, many literature annotation projects typically involve following steps:

- **Step** 1 : To take a sub-collection from the two databases, PubMed and/or PMC.
- **Step** 2 : To extract texts from each article, preprocessing them for annotation.
- **Step** 3 : To annotate the pre-processed texts

At Step 1, while many projects are interested in annotating only a sub-collection of the available articles from the two databases (Kim et al., 2008; Verspoor et al., 2012), recent progress of technology has enabled the production of annotation in a large scale, and there are a number of projects producing annotations to the entire PubMed and OAPMC (Björne et al., 2010; Wei et al., 2012; Rindflesch and Fiszman, 2003).

Given the high productivity of annotation in the area, it is generally recognized that interoperability of the annotations produced by different projects is important to leverage the progress of the community utilizing the resources. Standardization of the format or representation is one sort of such efforts, e.g., Linguistic Annotation Framework (LAF) (Ide and Romary, 2004), Open Linguistics[4], and BioC[5]. The work presented in this paper addresses another issue with regard to interoperability of annotation: *sharing reference texts and alignment of annotation*.

2 Proposal

A text is often thought as a sequence of characters, and an annotation is regarded as identifying a specific part (span) of the sequence, attaching a piece of information to it. It is, in fact, very similar to genome annotation: in both cases, a span to be annotated is specified by its begin and the end positions on its base sequence.

A bit more general case can be found in the geographic map annotation, e.g., *Google map*, where the target of annotation is a 2-dimensional "grid" space, which may be filled with geographical data, e.g., elevation, at its outset. Then, the map may be annotated with various types of information, e.g., restaurants. Note that, however, for the geographic map annotation to work properly, the coordinate system to facilitate the positioning in the grid space, e.g., *latitude* and *longitude*, needs to be standardized and shared by the users.

In the case of genome annotation, it is genome sequences whereon annotation instances

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.ncbi.nlm.nih.gov/pmc/
[3] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[4] http://linguistics.okfn.org/
[5] http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/BioC/

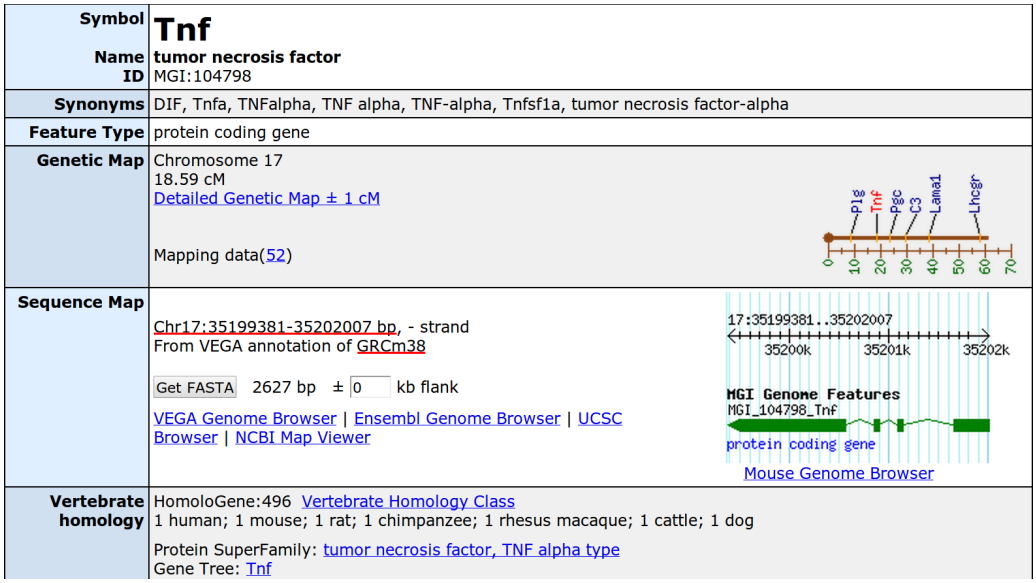


Figure 1: *Tnf* gene entry in the *MGI (Mouse Genome Informatics Genome)* database. The location of the gene, *Chr17:35,199,381-35,202,007*, and the specification of the reference sequence, *GRCm38*, is underlined red.

are based. Thus, it can be said that the genome sequence defines a coordinate system for genome annotation. Figure 1 shows an example of genome annotation which is taken from the *MGI (Mouse Genome Informatics)* database[6]. Roughly speaking, it shows that the location, *35,199,381 - 35,202,007*, on the 17'th chromosome of mouse is identified as the coding region of the protein, *Tumor necrosis factor* (which corresponds to *UniProt:P06804*). It is a part of the VEGA annotation which is made to *GRCm38*, a reference genome sequence of mouse provided by the *Genome Reference Consortium (GRC)*[7]. In the case, it can be said that the reference sequence defines the coordinate system for the specification of the gene coding region, and that without it the region specification loses its meaning.

The same may be applied to the literature annotation. Imagine that we want exchange a piece of annotation like as below:

```
([PMC:2626671, sec:1, span:14-22],
UniProt:P10820)
```

It may mean that "*the span between the 14'th and 22'nd characters of the section 1 of the literature, PMC:2626671, denotates UniProt:P10820 (which is the protein, Perforin-1)*". Note that the format of annotation is out of scope of this paper, and an annotation is simply represented as *n*-tuple throughout this paper. The piece of annotation, however, does not make a concrete sense without specifying the base text whereon the position specification holds its meaning.

Currently, the base texts are usually prepared by individual projects, making the position specification meaningful only when the project is specified:

```
(Project:A,
[PMC:2626671, sec:1, span:14-22],
UniProt:P10820)
```

Imagine that we know another piece of annotation produced by another project, *B*:

```
(Project:B,
[PMC:2626671, sec:1, span:14-22],
NP)
```

At a glance, it looks like the two projects, *A* and *B*, annotate the same span with the different labels, *UniProt:P10820* and *NP*, respectively. However, there is no evidence that the two spans are the same, as the base texts prepared by the two projects may be, at a high chance, different from each other. For example, one may include only *ASCII* characters, while the other includes *UTF-8* characters, and/or one may have inserted space characters for tokenization during preprocessing while the other may not. In such a situation, there is no direct way to compare or aggregate the annotations produced by different projects.

To remedy the situation, we propose to share the reference texts across annotation projects:

[6] `http://www.informatics.jax.org/`
[7] `http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/`

Organism	Literature
Genome sequencing	Text sequencing
Sequence alignment	Text alignment
Genome annotation	Text annotation

Table 1: Genome annotation vs. Text annotation

```
(Project:A,
[PMC:2626671, sec:0, span:17-25, ref:R],
UniProt:P10820)
(Project:B,
[PMC:2626671, sec:0, span:17-25, ref:R],
NP)
```

In the example, the two projects, *A* and *B*, annotate the same reference text, R, instead of producing the base texts themselves. Now, it is evident that the two pieces of annotation is made to the same span, i.e., the two projects share the same coordinate system that is defined by the reference texts. The two pieces of annotation are now immediately comparable and aggregatable: while the project *A* annotates the specific span as a protein, the project *B* annotates the same span as a noun phrase (NP).

3 PubAnnotation

To realize the scenario discussed in the previous section, PubAnnotation implements several functions as explained in the following sections.

3.1 Reference texts

In section 1, three typical steps for literature annotation in the area of life sciences is discussed. Among them, Step 2 is the main concern of this work, as the base texts of an annotation project is determined at the step. We call the process of producing texts from an article *text sequencing*, as it corresponds to the genome sequencing step of genome projects. Table 1 shows some corresponding steps of literature and genome annotation.

Often, text sequencing is thought as a trivial task, and its importance is neglected. It may be trivial if the target article is simple, e.g. PubMed abstracts. However, when it comes to full articles, e.g. PMC articles, it is not, as a full article often involves lots of non-linear structures, e.g., figures and tables. It is also not straightforward how to divide a full article into reasonably small parts for annotation, e.g., chapters, sections, or paragraphs. The choice of character encoding, e.g., *ASCII* or *UTF-8*, also has to be made at the step. Considering all the aspects, the text sequencing process needs to be fully automated for *reproducibility*.

PubAnnotation provides its own implementation of text sequencers for PubMed and PMC articles, which is freely available to the public, through persistent URLs. For example, the URL, `http://pubannotation.org/pmcdocs/2626671/divs/1`, refers to the text of the PMC document, PMC:2626671, in the second division[8]. As the texts are outputs of PubAnnotation sequencers, any piece of annotation to them needs to specify it to hold its meaning:

```
(Project:A,
[PMC:2626671, sec:0, span:17-25,
   ref:PubAnnotation],
UniProt:P10820).
```

As the PubAnnotation texts are universally accessible over the Web, the above piece of annotation remain valid as long as the Web is reachable.

The implementation detail of the text sequencers is out of scope of this paper. Instead, this paper focuses on how the annotations can be maintained valid while admitting the sequencers may be evolved over time.

3.2 Text conversion and alignment

While we suggest to share the base texts (so called reference texts) for annotation, we also admit that the base texts may be changed when necessary. For example, the output of the PubAnnotation sequencers includes Unicode characters to retain the content in the original article as much as possible. However, many annotation projects convert Unicode characters to equivalent *ASCII* sequences. One strong motivation for the conversion is that most NLP (natural language processing) tools, e.g., taggers and parsers, cannot properly handle Unicode characters. To address the need for text conversion, PubAnnotation features automatic text conversions and alignments.

Figure 2 illustrates an example. while the original text includes a Greek letter, β, one may want to convert it to an equivalent ASCII sequence, `beta`, to apply NLP tools, which may result in variation of the base texts. However, the annotation instances made to the varied texts have to be re-aligned to the reference texts to interoperability. PubAnnotation implements an automatic alignment of annotation using the Hunt-McIlroy's longest common subsequence (LCS) algorithm (Hunt and McIlroy, 1976) together with generalized LCS algorithm (Kim, 2013).

[8] *0*-oriented index

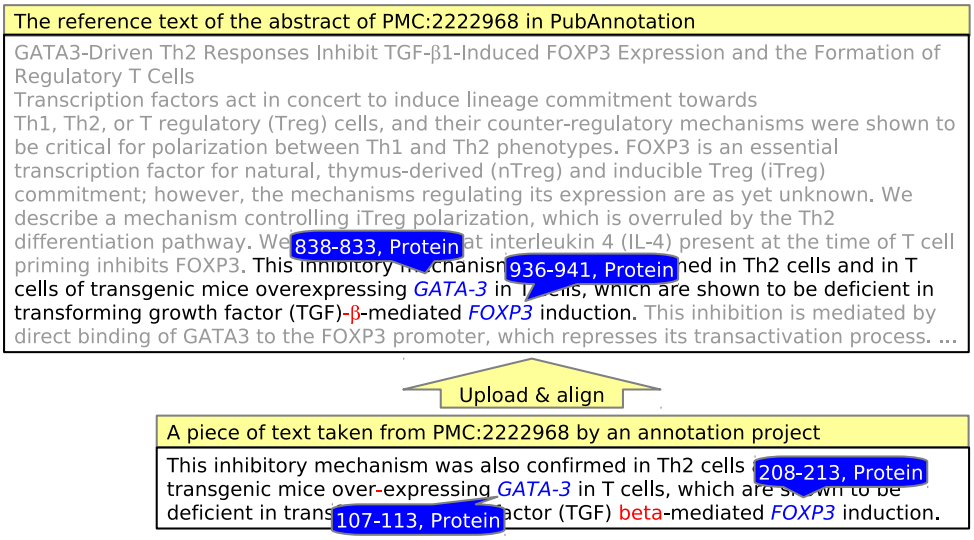


Figure 2: Illustration of annotation alignment to the reference text. The upper box represents the reference text provided by PubAnnotation. The lower box represents a piece of text taken by an annotation project. Note that the text is slightly different, e.g., β vs. *beta*, due to a preprocessing. The blue baloons illustrate that the two annotation instances, (107-113, Protein) and (208-213, Protein), are aligned to the reference text, by PubAnnotation.

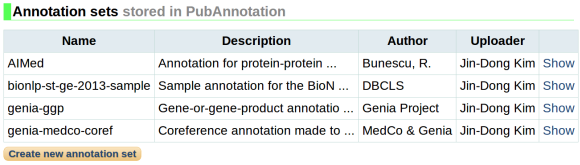


Figure 3: Screen-shot of *PubAnnotation*, listing all the annotation sets stored in it.

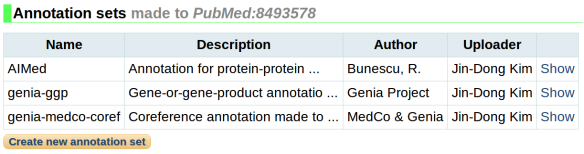


Figure 4: Screen-shot of *PubAnnotation*, listing the annotation sets made to the document: PubMed:8493578.

The function of automatic annotation alignment also enables seamless maintenance of annotation over the evolution or revision or sequencers, as discussed in the previous section.

4 Results

The PubAnnotation system is developed to provide a shareable repository of reference texts of life science literature and annotations to them. At the time of writing, the alpha-service is available at http://pubannotation.org. Figure 3 shows a screen-shot of *PubAnnotation*, listing the annotation projects stored on it. Note that all the annotations stored in PubAnnotation are aligned to the reference texts of PubAnnotation, and comparison of the annotations or collective analysis across them is immediately possible. Figure 4 shows all the annotation sets made to the document, *PubMed:8493578*. PubAnnotation is developed as an open-source project and downloadable under MIT license,

5 Conclusion

As an effort to improve the interoperability of literature annotation, the paper suggests to share the reference texts whereon annotations produced by different projects may be aligned. PubAnnotation is a public, open-sourced, repository system of reference texts and annotations, combined with web services which enable efficient maintenance of the annotations in various situation, e.g., text conversion and evolution of sequencers. We expect it to contribute to the reduction of the cost of the community sharing annotations.

Acknowledgment

This work is supported by the *Integrated Database Project* funded by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Jari Björne, Flip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Complex event extraction at PubMed scale. *Bioinformatics*, 26(12):382–390.

James W. Hunt and M. Douglas McIlroy. 1976. An Algorithm for Differential File Comparison. Technical Report 41, Bell Laboratories Computing Science.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Nat. Lang. Eng.*, 10(3-4):211–225.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from lterature. *BMC Bioinformatics*, 9(1):10.

Jin-Dong Kim. 2013. A Generalized LCS Algorithm and Its Application to Corpus Alignment. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1112–1116, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

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

Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A. Baumgartner Jr., Michael Bada, Martha Palmer, and Lawrence E. Hunter. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13:207.

Chih-Hsuan Wei, Bethany R. Harris, Donghui Li, Tanya Z. Berardini, Eva Huala, Hung-Yu Kao, and Zhiyong Lu. 2012. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database*, 2012:bas041.

Vocabulary Expansion by Semantic Extraction of Medical Terms

Maria Skeppstedt
DSV, Stockholm University
mariask@dsv.su.se

Magnus Ahltorp
Royal Institute of Technology
map@kth.se

Aron Henriksson
DSV, Stockholm University
aronhen@dsv.su.se

Abstract

Automatic methods for vocabulary expansion are valuable in supporting the development of terminological resources. Here, we evaluate two methods based on distributional semantics for extracting terms that belong to a certain semantic category. In a list of 1000 terms extracted from a corpus of Swedish medical text, the best method obtains a recall of 0.53 and 0.88, respectively, for identifying 90 terms that are known to belong to the semantic categories Medical Finding and Pharmaceutical Drug.

1 Introduction

High-coverage terminologies are important for medical text processing systems, such as named entity recognizers and information extractors. Manual terminology development is, however, expensive and time-consuming; it also runs the risk of resulting in insufficiently extensive terminologies and a subsequent negative impact on the recall of systems in which these are used. Methods that can support this process in various ways are thus very valuable.

Given the availability of a large corpus, methods based on distributional semantics – i.e. methods that exploit term co-occurrence patterns – make it possible to determine, in an unsupervised fashion, which terms are semantically related and to what extent. Several studies have demonstrated the potential of these methods in the (bio)medical domain (Cohen and Widdows, 2009), also with clinical corpora for the purpose of semi-automatic medical vocabulary development (Henriksson et al., 2012) and query expansion (Zeng et al., 2012).

Previous applications of distributional semantics for terminology development support and similar tasks have either focused on the extraction of very closely related terms, e.g. synonyms (Landauer and Dumais, 1997; Henriksson et al., 2013), or used features derived from such methods to train named entity recognition systems (Sahlgren and Cöster, 2004; Jonnalagadda et al., 2012). Here, we aim to study more closely the potential of using distributional semantics to extract terms that belong to a specific semantic category of medical terms, which will hopefully contribute to the areas of semi-automatic terminology development and unsupervised feature extraction.

2 Background

Methods for automatic vocabulary extraction can be divided into two main types, depending on whether or not there already exists a terminology (or a set of seed words belonging to predefined semantic categories). With the availability of a terminology in the target domain, as is the case in this study, vocabulary extraction can be seen as a classification task, determining whether an unknown word belongs to a certain semantic category. If there does not yet exist a suitable resource, however, a clustering approach needs to be taken, where clusters constitute candidates for semantic categories. In either case, the vocabulary extraction is based on finding patterns of contexts in which words typically occur (Biemann, 2005).

Semantic (word) spaces, derived from a corpus, represent such context patterns in the form of word co-occurrence information. This representation has been used both for creating clusters of semantically related words (Song et al., 2007) and for determining whether unknown words belong to predefined semantic categories (Widdows, 2003; Curran, 2005). In this study, we use a computationally light-weight version of the semantic space representation called *random indexing* (Kanerva et al., 2000; Karlgren and Sahlgren, 2001; Sahlgren, 2005). Instead of reducing the dimensionality of a word-by-word (or word-by-context) matrix to

make it computationally tractable (which is the approach taken for creating many other types of semantic spaces), a matrix with a smaller dimensionality is created from the beginning. Each word in the corpus is assigned a unique representation in the form of an *index vector* with a dimensionality that is much smaller than the number of unique terms in the corpus. The near-orthogonal index vectors are created by randomly generating very sparse vectors, in which most of the elements are set to 0, while a few (1–2%), randomly selected, elements are set to either $+1$ or -1. Each word is also assigned a *context vector* with the same dimensionality as the index vector, in which all elements are initially set to 0. For every occurrence of a word in the corpus, its context vector is updated by adding the index vectors of the words in the context window (the surrounding words). Different semantic relations can be modelled by varying the size of the context window (Sahlgren, 2006). The resulting semantic space consists of the context vectors, between which, e.g., the cosine similarity can be computed to determine the semantic distance between words.

3 Materials and Methods

The proposed approach essentially requires two resources: a large corpus of medical text and a number of seed terms that belong to the semantic category of interest. To allow the method(s) to be evaluated automatically, additional terms that are known to belong to the same semantic category are also needed. Here, a corpus of Swedish medical text and subsets of the Swedish version of the medical vocabulary MeSH were used.

3.1 Semantic Spaces of Medical Text

Semantic spaces were induced from a Swedish medical corpus: Läkartidningen, which is the Journal of the Swedish Medical Association (Kokkinakis, 2012) and contains articles on, for instance, new scientific findings in medicine, pharmaceutical studies and health-economic evaluations. Editions from the years 1996–2005 were used, as these have been made available for research, albeit with the sentences given in a random order. The corpus was preprocessed by (white-space) tokenising and lower-casing the text. Since the sentence order is scrambled, a document break was inserted between sentences to ensure that co-occurrence information is not collected across sentences in the construction of the semantic spaces. The corpus was not lemmatised, as inflected forms of medical terms may also be relevant candidates for vocabulary expansion. The corpus contains 21 447 900 tokens and 444 601 unique terms.

Random indexing was applied to induce 1000-dimensional semantic spaces[1] from variants of this corpus. The semantic spaces were evaluated in two steps: (1) in a development phase, where context window size was optimised separately for each of the two semantic categories (Medical Finding and Pharmaceutical Drug) and for each of the two proposed methods, and (2) in a final evaluation phase, where the best-performing semantic spaces, in terms of recall, were evaluated on unseen data. The context window sizes 1+1, 2+2, 4+4 and 50+50 were evaluated in the development phase. The 50+50 window size is, in effect, a sentence-level context definition since the sentence delimiters ensure that context information from adjacent sentences is ignored.

3.2 Semantic Term Extraction

Two computationally efficient methods for vocabulary expansion using random indexing were devised and evaluated: *Term Replacement (TermRep)* and *Cosine Addition (CosAdd)*.

In the first method, *TermRep*, the corpus was modified before the semantic spaces were created. All occurrences of a set of seed terms that belong to a given semantic category were replaced by a common string denoting that category. This can be seen as an aggressive form of term normalisation and entails that each semantic category is assigned a single context vector, which is populated with the index vectors of terms that co-occur with all lexical instantiations of that semantic category. The string that represents the semantic category of interest was then given as a query term to the semantic space, resulting in a ranked list of distributionally similar terms, presumably some of which belong to the same semantic category.

In the second method, *CosAdd*, the semantic spaces were created with the unmodified corpus. Each term in the set of seed terms was instead used as a query term, resulting in one ranked list per seed term, containing the cosine similarity between this seed term and every other word in the

[1] 10 non-zero elements (i.e., 1%) were assigned to the index vectors. When populating the context vectors, increasingly less weight was assigned to index vectors as the distance from the target term increases.

corpus. The ranked lists of the seed terms were then merged into a single ranked list per semantic category. The merge was performed by summing the cosine similarity scores.

A certain number of observations of a term is required for its context vector to be accurately positioned. Words occurring fewer than 50 times were therefore not included as seed terms; they were also excluded from the lists of candidate terms.

3.3 Medical Terminology and Evaluation

The medical terminology was here employed for two purposes: (1) as a set of seed terms for a given semantic category and (2) as a reference standard for evaluating the two proposed methods.

The Swedish version of MeSH[2] (Karolinska Institutet, 2012), a controlled vocabulary for indexing life science literature, was here used for these purposes. For the semantic category Medical Finding, terms that belong to the Swedish MeSH categories *Disease or syndrome* and *Sign or symptom*[3] were used; for the semantic category Pharmaceutical Drug, the MeSH category *Pharmacologic substance* was used.

MeSH terms occurring fewer than 50 times were excluded as seed terms (as mentioned above), as well as reference standard terms. Multiword terms were also excluded, as current models of distributional semantics perform better on unigram terms (Henriksson et al., 2013). When rare and multiword terms had been removed, 309 terms that belong to Medical Finding and 181 terms that belong to Pharmaceutical Drug remained. In order to enable a fairer comparison between the two semantic categories, 181 Medical Finding terms – identical to the number of Pharmaceutical Drug terms – were randomly selected.

The terms used in the evaluation for each semantic category were divided into two stratified, equally large groups, a *development set* and an *evaluation set*, in which the strata consisted of terms with similar frequencies in the corpus. In the development phase, the terms in the development set were used for optimising context window size. In the evaluation phase, all terms were used: the terms in the development set were treated as seed terms, which, in a real-world scenario, would be known and already included in the terminology; the terms in the evaluation set were ones that, in a real-world scenario, we would like to add to the terminology.

The performance using different window sizes was measured using 10-fold cross-validation on the data in the *development set*. The 91 terms that belong to Medical Finding and the 91 terms that belong to Pharmaceutical Drug were divided into ten folds. That is, for each fold, approximately 82 terms were used as query terms – or, in the *TermRep* case, replaced by a common identifier in the corpus – and approximately 9 terms were expected to be retrieved, effectively making up the reference standard. Recall was measured as the proportion of expected terms that were found in a list of retrieved terms. Recall at different cut-off values (from 50 to 1000, with a step size of 50) were calculated. The semantic spaces with the highest average recall values were selected and used in the evaluation phase. This means that the semantic spaces were not optimised for a specific cut-off value, rendering the cut-off value a flexible parameter in the final evaluation.

In the evaluation phase, the primary evaluation was conducted in the form of a fully automatic evaluation of recall against the *evaluation set*. To determine to what extent retrieved terms belong to the expected semantic category, despite not being present in the reference standard, a semi-automatic evaluation of precision among the 500 top-ranked terms was also performed. Retrieved terms classified as Finding or Drug in MeSH or FASS (2012) were automatically classified as correct or incorrect (assuming that a known Finding can never be a Drug and vice versa). The remaining terms were manually classified by a single annotator as belonging to the category or not.

4 Results

Averaging the recall measurements for the 20 cut-off values yielded the results shown in Table 1. There were no large differences between window sizes, but the best recall (for both methods) was obtained with a context window of 2+2 for Medical Finding and 1+1 for Pharmaceutical Drug. Semantic spaces induced with these window sizes were therefore used in the final evaluation.

The ability of the two methods to extract the expected terms in the evaluation set is shown in Figure 1. For Medical Finding there was no large difference between the two methods, whereas *Cosine*

[2]**ME**dical **S**ubject **H**eadings: `http://www.nlm.nih.gov/mesh`

[3]As there is a rather fine distinction between these two subcategories, they were merged into a single category.

Window Size	1+1	2+2	4+4	50+50
	Medical Finding			
CosAdd	0.372	**0.389**	0.384	0.382
TermRep	0.357	**0.368**	0.361	0.360
	Pharmaceutical Drug			
CosAdd	**0.567**	0.516	0.502	0.501
TermRep	**0.409**	0.386	0.375	0.371

Table 1: Average recall values over 20 different cut-offs (top 50 – top 1000) on development data.

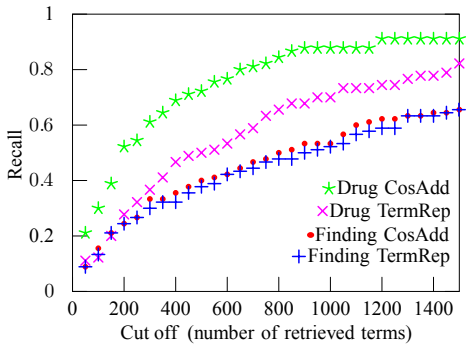


Figure 1: Recall values for different cut-offs

Addition outperformed *Terminology Replacement* for Pharmaceutical Drug. Both methods obtained better recall for extracting Drug terms than Finding terms. The overlap of retrieved terms for the two methods was 83% for Finding and 76% for Drug (top 1000). For the *CosAdd* method, precision was also evaluated, with better results for Finding than for Drug (0.80 vs. 0.64 for top 50 and 0.68 vs. 0.47 for top 100, Figure 2).

5 Discussion

Two computationally light-weight methods for automatic vocabulary expansion have been studied.

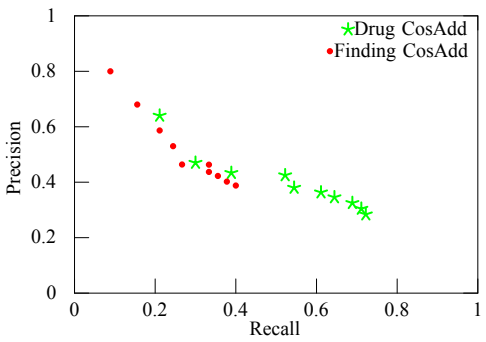


Figure 2: Precision (partially based on manual classification) vs. recall (automatically measured against the reference standard), cut-off 50–500.

Seed terms were modelled as if they would form two separate clusters in the semantic space: one for Medical Finding and one for Pharmaceutical Drug. When applying the replacement method, we are in effect searching for new words that are close to a weighted centroid of the cluster. The weighting emerges from the fact that the effect of each seed term on the resulting centroid context vector is directly proportional to the frequency of the seed term in the corpus. This makes the method vulnerable to frequent seed terms that are atypical for the semantic category, which might explain the lower results with this method for Pharmaceutical Drug, as, for instance, *alcohol* was the second most frequent seed term. With the addition method, on the other hand, each seed term is given equal weight, and new words are deemed equally typical to the semantic category irrespective of the frequency of the seed term to which they are close. This means that employing a low frequency threshold for which seed terms to include might drastically lower the results, as there is a weak statistical foundation for the position of the context vectors of the many low-frequent terms.

6 Conclusion and Future Work

The best performing method was able to extract 53% of the 90 expected Medical Findings and 88% of the 90 expected Pharmaceutical Drugs among the top 1000 retrieved terms, showing its potential as a useful component in a semi-automatic vocabulary expansion process. Future work should, however, include a comparison between the approaches evaluated here and previous approaches, for their ability to retrieve expected terms and also for their computational efficiency.

Moreover, modelling a MeSH category as one cluster in the created semantic space is most likely an over-simplification. There might be a number of sub-clusters within each of the two categories Finding and Drug – sub-clusters that are positioned at large distances from each other in the semantic space. Words not part of these sub-clusters, but close to two or more clusters, will then receive a high ranking with the methods applied here, even though they ought to be ranked lower than words close to the centroids of the sub-clusters. As the next step, we will therefore attempt to cluster the seed terms into sub-clusters and apply the distance measures of this study to rank the similarity of unknown words to these sub-clusters.

Acknowledgments

We are very grateful to Rafal Rzepka and Shiho Kitajima for their valuable feedback on the study. We would also like to thank the three reviewers for many good comments.

This work was partly supported by the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (ref. no. IIS11-0053) at Stockholm University, Sweden.

References

Chris Biemann. 2005. Ontology learning from text: A survey of methods. In Alexander Mehler, editor, *Themenschwerpunkt Korpuslinguistik, GLDV-Journal for Computational Linguistics and Language Technology*. Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV).

Trevor Cohen and Dominic Widdows. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390 – 405.

James R. Curran. 2005. Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 26–33, Stroudsburg, PA, USA. Association for Computational Linguistics.

FASS. 2012. Fass.se. http://www.fass.se, Accessed 2012-08-27 08-27.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, Vidas Daudaravičius, and Martin Hassel. 2012. Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models. In *Proceedings of the 5th International Symposium on Semantic Mining in Biomedicine (SMBM)*.

Aron Henriksson, Maria Skeppstedt, Maria Kvist, Martin Duneld, and Mike Conway. 2013. Corpus-Driven Terminology Development: Populating Swedish SNOMED CT with Synonyms Extracted from Electronic Health Records. In *Proceedings of BioNLP*. Association for Computational Linguistics.

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

Pentti Kanerva, Jan Kristofersson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of 22nd Annual Conference of the Cognitive Science Society*, page 1036.

Jussi Karlgren and Magnus Sahlgren. 2001. From words to understanding. *Foundations of Real-World Intelligence*, pages 294–308.

Karolinska Institutet. 2012. Hur man använder den svenska MeSHen (In Swedish, translated as: How to use the Swedish MeSH). http://mesh.kib.ki.se/swemesh/manual_se.html. Accessed 2012-03-10.

Dimitrios Kokkinakis. 2012. The journal of the Swedish medical association - a corpus resource for biomedical text mining in Swedish. In *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop. Turkey*.

Thomas K Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Magnus Sahlgren and Rickard Cöster. 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 487. Association for Computational Linguistics.

Magnus Sahlgren. 2005. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5.

Magnus Sahlgren. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, PhD thesis, Stockholm University.

Dawei Song, Guihong Cao, Peter D. Bruza, and Raymond Lau. 2007. Concept induction via fuzzy c-means clustering in a high-dimensional semantic space. In J. Valente de Oliveira and W. Pedrycz, editors, *Advances in Fuzzy Clustering and its Applications*, pages 393–403. John Wiley & Sons, Chichester.

Dominic Widdows. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 197–204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Qing T Zeng, Doug Redd, Thomas Rindflesch, and Jonathan Nebeker. 2012. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. In *AMIA Annual Symposium Proceedings*, pages 1050–1059.

Impact of real data from electronic health records on the classification of diagnostic terms

Alicia Pérez
IXA Taldea
(UPV-EHU)

Koldo Gojenola
IXA Taldea
(UPV-EHU)

Maite Oronoz
IXA Taldea
(UPV-EHU)

Arantza Casillas
IXA Taldea
(UPV-EHU)

Abstract

This work tackles Electronic Health Record (EHR) classification according to their Diagnostic Terms (DTs) following the standard International Classification of Diseases-Clinical Modification (ICD-9-CM). To do so, we explore text mining relying on a wide variety of data from both standard catalogues, such as the ICD-9-CM and SNOMED-CT; and, what it was proven even more effective, real data sources, such as EHRs.

The models we put forward to deal with this problem are Finite-State Transducers (FSTs). The aim behind FSTs would be not only to accept exact terms in the ICD-9-CM but also alternative variants. To be precise, a series of FSTs were defined to carry out a soft-matching process between DTs written in natural language and those in the standard form as in the ICD-9-CM catalogue.

1 Introduction

The Clinical Documentation Service of the Galdakao-Usansolo Hospital (a hospital attached to the Spanish Ministry of Health, Social Services and Equality) is interested on automatising the classification of Electronic Health Records (EHRs). EHRs include several fields such as: a description of the patient's details, antecedents, procedures and methods of administration of medicines, and Diagnostic Terms (DTs). It is the DTs that serve as the classification key to classify EHRs according to the World Health Organisation's 9th Revision of the International Classification of Diseases - Clinical Modification (ICD-9-CM)[1]. The goal of this work is to develop a system to automatically classify DTs in an attempt to alleviate the work load by the Clinical Documentation Service but never at the expense of precision. This task presents the following challenges:

1. Natural language in EHRs vs. medical jargon in ICD-9-CM
2. Large-scale classification problem: including more than 14×10^3 different classes
3. Working towards a 100% precision

1.1 Related work

A large number of sophisticated machine learning algorithms have been applied to the task of DT classification. Ferrao et al. (2012) used a commercial system based on either Naive-Bayes or decision trees to tackle multi-label classification of EHRs restricted to the Internal Medicine department.

The top systems in the 2007 Computational Medicine Challenge have benefited from incorporating domain knowledge of free-text clinical notes, such as negation, synonymy and hyperonymy, either as hand-crafted rules in a symbolic approach, or as carefully engineered features in a machine learning component: (Goldstein et al., 2007; Crammer et al., 2007; Aronson et al., 2007; Patrick et al., 2007). Yet, this shared task involved the assignment of ICD-codes to radiology reports written in English from a reduced set of 45 codes (Pestian et al., 2007). By contrast, we focus on the entire scope of the ICD-9-CM catalogue.

Most of the systems described in the literature were developed for English. Looking at other languages, Metais et al. (2007) reported a system to classify medical reports in French.

[1]The reader might be aware of the fact that for English other codification systems (such as ICD-10) are also reported in the literature, nevertheless, it is the ICD-9-CM the one being currently used by the Spanish Health System even though it is foreseen to move to ICD-10 in the near future.

2 Methods: Finite-State Transducers

Finite-State Automata (FSA) serve to the purpose of recognising regular grammars (Chomsky, 1959). A grammar is used to either generate or parse the strings accepted in the language recognised by the FSA. In our medical domain the DTs in the ICD-9-CM catalogue represent the set of acceptable strings within a formal language with a particular syntax. Thus, inferring the grammar underlying the DT domain would help to assess whether a given string could be considered or not appropriately expressed in that language.

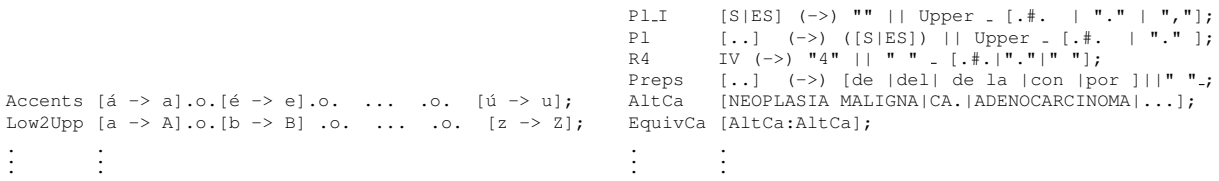
Finite-State Transducers (FSTs) are an extension of FSAs that encompasses two languages: input and output. FSTs serve to analyse an input string and associate an output string (in case that the input is acceptable in the source language). That is, FSTs serve to map from one language to the other. The nature of the FSTs does not allow to accept any string out of the language, and this property strives towards a high precision.

2.1 Implementation

In brief, the system is designed as a composition of three FSTs: lexicon, normalisation and generation. The FSTs were next integrated on a priority union basis. This operation allows a wide search while it tries to stick as possible to the input. Besides, it rejects some strings, meaning that it reveals ill-formed DTs. All the FSTs as well as their operations were implemented through Foma (Hulden, 2009). Foma is a freely available toolkit that allows to build finite-state transducers and also includes efficient parsing functions. Besides, it supports imports from, and exports to, other toolkits, such as Xerox's XFST (Beesley and Karttunen, 2003), AT&T (Mohri et al., 2003) and OpenFST (Riley et al., 2009). Next we provide some details of each FST:

1. **FST-Lexicon:** it compiles the reference collection of allowed (DT, ICD-code) pairs, that is, the lexicon of the application. This FST is automatically built by Foma from the set of pairs allowed. The data-sets involved in the lexical model came from two sources:
   - ICD-9-CM: consists of more than $14,435$ different (DT, ICD-code) pairs not restricted to a single clinical domain.
   - EHRs in Spanish: a set of more than 28,000 (DT, ICD-code) pairs with DTs written by doctors and coded by experts in EHRs that allows supervised classification.
2. **FST-Normalisation:** it carries out elementary pre-processing operations. The goal is to get all the inputs re-cased, to get rid of written accents and other punctuation marks that are considered as noisy. This FST was built from rules and compiled as an FST by Foma. An example of the rules underlying this FST is given in Figure 1a.
3. **FST-Generation:** it allows to generalise the reference lexicon by means of synonyms, acronyms, etc. As a result, it allows to generate new alternatives for the DTs. This FSTs implements rules to check punctuation marks, to allow number variation (to create singular and plural forms for a given DT in the reference), the omission and equivalence of some prepositions, either expand abbreviations, synonyms of the reference according to SNOMED-CT, optional replacement in a given context, composition, union, projection,etc. For exemplification purposes, some of these rules are shown in a very simplified manner in Figure 1b.

Let us show in an example the procedure by which the system makes it possible the automatic assignment of the correct ICD-code, 185, to the DT "Ca. prostata" used in an EHR. In the ICD-9-CM the term encoded with 185 is "Neoplasia maligna de la próstata". Hence, an exact lookup operation would have been unproductive. Nevertheless, the soft-matching operations implemented through the proposed FST are able to find the required term, and accordingly, provide the corresponding ICD-code. As a first step, both terms (the DT and the one in the ICD list) are normalised by the FST-Normalisation that was defined from the set of rules denoted as `Accents` and `Low2Upp` (see Figure 1a). The normalisation step yields "CA. PROSTATA" and "NEOPLASIA MALIGNA DE LA PROSTATA". After that, the FST-Generation proceeds with the generation of several alternatives: the `AltCa` rule enables the equivalence of several alternatives, such as "CA." and "NEOPLASIA MALIGNA". Hence, this enables to parse "CA. PROSTATA" as "NEOPLASIA MALIGNA PROSTATA". Finally, the `Preps` rule adds the prepositions, leading to the standard



(a) Normalisation

(b) Generation

Figure 1: Rules underlying the FSTs involved: FST-Normalisation and FST-generation

term in the ICD list "NEOPLASIA MALIGNA DE LA PROSTATA" from the DT in the EHR "CA. PROSTATA".

The FSTs were arranged with a priority union in such a way that each FST contributed with additional capabilities to the previous one. The transducers were composed in such a way that the most simple transducer was looked-up first and the one allowing the higher variability last. That is, a priority union is applied to compose the different transducers.

3 Experimental results

For this task it is preferred to get accurate results with high precision even at the expense of low coverage. Hence, the system allows rejections whenever the input DT does not match any of the alternatives allowed in the language accepted by the FST. That is, all the instances that did not soft-match a DT in the FST are left unclassified and this is why we are not referring to our system as a fully automatic classification system but as a classification support system, instead.

Accordingly, for a given DT there are three possible outcomes:

Reject: the DT was not assigned any code by the system because the input DT did not soft-match any of the accepted alternatives in the FST. That is, there was not any path in the transducer accepting the source string.

Miss: the DT was assigned a code by the system that did not match the manually assigned ICD-code.

Hit: the DT was assigned a code that matched the one in the reference.

The performance of the FST, shown in Table 1, was assessed using a 5-fold cross validation on the EHR set of 28,000 (DT, ICD-code) pairs, while including also the ICD-9-CM set to feed the FST-Lexicon.

In order to make clear the relevance of both the nature of the seed lexicon and the generation operation, we made a baseline experiment: the lexicon consisted only of the standard ICD-9-CM set of pairs and while normalisation operation was allowed, we did not allow for any generation. Through this *baseline* we meant to measure the number of DTs written by doctors nearly as in the standard ICD-9-CM. Although the ICD-9-CM is composed of 14,435 different pairs, the number of hits achieved was 7.1%. Moreover, allowing next the generation operation on the same lexicon, the hits represent the 8.1%, the rejections the 89.0% and the misses the 2.9%. Comparing this baseline with the results in Table 1, the conclusion drawn is that the aid of real EHRs seems to be of much benefit in what comes to feeding the lexicon of the FST.

Evaluation	Rejections	Misses	Hits
automatic	12.0%	1.2%	86.8%

Table 1: Performance of the FST.

3.1 Impact of real data on performance

Having incorporated EHRs to the allowed lexicon provided excellent results with respect to the baseline. Hence, it seemed of interest to quantitatively assess the impact of including more and more instances from EHRs, which is, precisely, one of the hubs of this paper.

The aim is to learn a regression model that would predict the effect of adding further data on the coverage. To do so, more and more instances from EHRs were progressively added to the lexicon and the improvements in terms of coverage were evaluated. A polynomial regression on the evaluation data was carried out showing the following approximated relation:

$$y \approx f(x) = a_2 \cdot x^2 + a_1 \cdot x + a_0 \qquad (1)$$

being:

x the size of the (DT, ICD-code) pairs from EHRs used to feed the FST-Lexicon, presented in logarithmic scale.

y the number of rejections provided by the FST, expressed as a percentage.

to be precise:

$$x = ln(|\mathcal{C}|) \quad (2)$$

$$y = \frac{|\mathcal{R}| * 100}{|\mathcal{R}|} \quad (3)$$

On this basis, a quadratic polynomial predictive model presented in eq. (1) was derived with the following coefficients:

$$a_2 = 1.57 \qquad a_1 = -37.5 \qquad a_0 = 226 \quad (4)$$

These results, represented in Figure 2, show that even a small corpus would represent a leverage to gain on coverage for similar tasks.

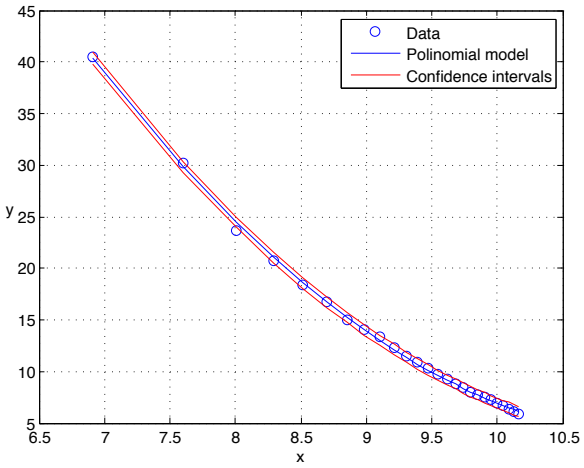


Figure 2: The number of rejections as a percentage (in the ordinate) with respect to the size of the corpus in logarithmic scale (abscissa). Experimental results are represented as circles. The quadratic polynomial function proposed in eq. (1) is represented together with its confidence interval by the curve and its upper and lower bounds.

The experimental results show that the corpus plays a core role on the performance of the system. While the standard ICD list showed to be of help, significantly better results were obtained extracting the lexicon from previously classified DTs written in EHRs. The impact of adding more and more DTs from previous EHRs to the corpus has shown to reduce the number of unclassified DTs in a logarithmic basis. Moreover, as a side effect the precision was also improved.

4 Concluding remarks and future work

In this work we present a system to classify diagnostic terms in Spanish according to the ICD-9-CM standard. The approach was based on the representation of a corpus of (DT, ICD-code) pairs in terms of FSTs that would parse an input DT into an output ICD-code.

The experimental results showed that the corpus played a core role on the performance of the system. The role played by the corpus opens another line of research: possibly lower amounts of data could be used with similar performance making use of adaptive models for different user-profiles (writing styles, use of abbreviations, etc.).

To sum up, the contribution of this paper are:

1. Large-scale and high precision automatic DT classification: the main contribution of this work is a high precision automatic classification of DTs in EHRs according to the ICD-9-CM reference. We propose the use of the FST framework, that allows not only to do an exact lookup but also a soft-matching within the lexicon or a set of positive samples.

2. Quantification of the benefits of real data: we propose the use of previously classified corpus in order to enhance the matching process adding DTs written differently to the standard.

3. Development of medical resources in Spanish: to the authors' knowledge this is the first attempt using all the codes in the ICD list in Spanish and rule-based pattern recognition approach. In addition, we contributed with an underlying process of acquisition and also with a pre-processing of valuable lexical resources within the medical domain in Spanish.

Future work will focus on those DTs that were rejected by the system (and thus, left unclassified) in an attempt to gain coverage. Together with FSTs, other strategies, such as support vector machines shall be explored. While this work was presented as an automatic classification approach, since the goal is to arise a 100% precision, it seems of interest to explore the unclassified DTs through interactive pattern recognition approaches (Toselli et al., 2011). This is can also be achieved through FSTs, since they were proven efficient in computer-aided tasks.

Acknowledgments

We would like to thank the Hospital Galdakao-Usansolo for their contributions and support, in particular to Javier Yetano, responsible of the Clinical Documentation Service.

This work was supported by the Department of Industry of the Basque Government (IT344-10, S-PE12UN114), the University of the Basque Country (GIU09/19), the Spanish Ministry of Science and Innovation (TIN2010-20218, TIN2012-38584-C06-02).

References

[Aronson et al.2007] A.R. Aronson, O. Bodenreider, D. Demner-Fushman, K.W. Fung, V.K. Lee, J.G. Mork, A. Neveol, L. Peters, and W.J. Rogers. 2007. From indexing the biomedical literature to coding clinical text: Experience with MTI and machine learning approaches. In *Proceedings of the Workshop on BioNLP*, pages 105–112.

[Beesley and Karttunen2003] Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications,.

[Benesch et al.1997] C Benesch, DM Witter, AL Wilder, PW Duncan, GP Samsa, and DB Matchar. 1997. Inaccuracy of the international classification of diseases (icd-9-cm) in identifying the diagnosis of ischemic cerebrovascular disease. *Neurology*, 49(3):660–664.

[Chomsky1959] Noam Chomsky. 1959. On certain properties of formal grammars. *Information and Control*, 2(2):137—167.

[Crammer et al.2007] K. Crammer, M. Dredze, K. Ganchev, P.P. Talukdar, and S. Carroll. 2007. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP*, pages 129–136.

[Ferrao et al.2012] J.C. Ferrao, M.D. Oliveira, F. Janela, and H.M.G. Martins. 2012. Clinical coding support based on structured data stored in electronic health records. In *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, pages 790–797.

[Goldstein et al.2007] I. Goldstein, A. Arzumtsyan, and O. Uzuner. 2007. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. In *Proceedings of the AMIA Annual Symposium*, pages 279–283.

[Hulden2009] Mans Hulden. 2009. Foma: a Finite-State Compiler and Library. In *European Association for Computational Linguistics*, pages 29–32. The Association for Computational Linguistics (ACL).

[Lita et al.2008] Lucian Vlad Lita, Shipeng Yu, Radu Stefan Niculescu, and Jinbo Bi. 2008. Large scale diagnostic code classification for medical patient records. In *Third International Joint Conference on Natural Language (IJCNLP)*, pages 877–882, Hyderabad, India, January. The Association for Computer Linguistics.

[Metais et al.2007] Elisabeth Metais, Didier Nakache, and Jean-François Timsit. 2007. Automatic classification of medical reports, the cirea project. In *Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics*, pages 354–359.

[Mohri et al.2003] Mehryar Mohri, Fernando C. N. Pereira, and Michael D. Riley. 2003. AT&T FSM LibraryTM – Finite-State Machine Library.

[Patrick et al.2007] J. Patrick, Y. Zhang, and Y.Wang. 2007. Evaluating feature types for encoding clinical notes. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 218–225.

[Pestian et al.2007] John P. Pestian, Chris Brew, Pawel Matykiewicz, D. J Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic, June. Association for Computational Linguistics.

[Riley et al.2009] Michael Riley, Cyril Allauzen, and Martin Jansche. 2009. OpenFST: An open-source, weighted finite-state transducer library and its applications to speech and language. In *Proceedings of Human Language Technologies. Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 9–10. Association for Computational Linguistics.

[Toselli et al.2011] Alejandro H. Toselli, Enrique Vidal, and Francisco Casacuberta. 2011. *Multimodal Interactive Pattern Recognition and Applications*. Springer.

Comparing Social Media and Search Activity as Social Sensors for the Detection of Influenza

Mizuki Morita
National Institute of Advanced Industrial Science and Technology (AIST)
mori-ta.mizuki@aist.go.jp

Sachiko Maskawa
Photonic System Solutions
sachiko.maskawa@gmail.com

Eiji Aramaki
Kyoto University/PRESTO
ei-ji.aramaki@gmail.com

Abstract

Detecting the incidence and prevalence of infectious diseases is important because these diseases affect many people, raise the cost of healthcare, and, in some cases, can lead to a great many deaths. Recently, it has been shown that people's online activity can be applied to detecting the prevalence of influenza. In this study, we compare the characteristics of two kinds of online activities, social media and search activity, as the social sensors capable of supplying information on seasonal epidemics and an unexpected pandemic of influenza. Although both approaches showed quite high performance for the seasonal epidemics, they showed poor performance for the unexpected influenza pandemic. The social media-based approach particularly over-responded to the influenza pandemic.

1 Introduction

Influenza has persisted as a major worldwide public health concern. Although it was discovered early in the last century that a virus causes influenza (Yamanouchi et al., 1919; Smith et al., 1933), influenza has persisted in threatening people and has led to the deaths of a huge number of people around the world (World Health Organization, 2003). Seasonal influenza epidemics typically occur in winter across temperate regions of the world. Although unexpected influenza pandemics rarely occur, we experienced them three times in the 20th century: the Spanish flu, Asian flu, and Hong Kong flu (World Health Organization, 2009).

The routes of transmission of influenza are known. Therefore, effective ways exist to prevent transmission of influenza, such as vaccination; keeping hands clean; avoiding touching the mouth, nose, and eyes; and using a face mask (World Health Organization, 2010). For early detection of the onsets of influenza epidemics and pandemics, some countries have adopted public surveillance agencies such as the Center for Disease Control and Prevention (CDC) in the US, the European Centre for Disease Prevention and Control (ECDC) in the EU, and the Infectious Disease Surveillance Center (IDSC) in Japan (note that the early detection of infectious diseases is an important task, but it is not the organizations' sole reason for existence). A major concern is that reports from these agencies typically have a time lag of 1–2 weeks.

Recently, by aiming at earlier detection of influenza onset, internet-based approaches have adopted access logs of health-related websites (Johnson et al., 2004), search queries to the popular search engine Yahoo! (Polgreen et al., 2008), search queries to a medical website (Hulth et al., 2009), search queries to Google (Ginsberg et al., 2009), and posts on the microblogging site Twitter (Aramaki et al., 2011; Signorini et al., 2011). People's behavior on the internet can be divided into two camps: communicating with others, such as through posting on social media (e.g., Twitter, Facebook, and LinkedIn), and searching for themselves, such as querying search sites (e.g., Google, Yahoo! and Microsoft Bing).

In this study, we have attempted to characterize both social media-based and search activity-based approaches to detecting influenza in respective case of seasonal epidemic and unexpected pandemic influenza prevalence.

2 Methods

We implemented an approach to estimate the number of influenza patients based on Twitter posts in Japanese (Aramaki et al., 2011). We also used results from a Google search query-based approach (Ginsberg et al., 2009). We evaluated the performance of these approaches against public surveillance data in Japan.

Here, the annual influenza season in Japan typically starts from November through December and subsides sometime in April or May, although the trends of the number of patients and the extent of epidemics vary from year to year. In 2009, an extremely important public health issue facing the world was influenza A (H1N1), also known as "swine flu," the first influenza pandemic of the 21st century. In Japan, the influenza A(H1N1)pdm09 viruses were first detected in three returning travelers from abroad on May 9, 2009. After several hundred patients with influenza A (H1N1) had been found, the Japanese Ministry of Health, Labour and Welfare (MHLW) presented a perspective that the number of newly infected patients was decreasing in late May.

Twitter data

We collected Twitter posts in Japanese during November 2008 – July 2009 and November 2012 - June 2013 via the Twitter API (`https://dev.twitter.com/`). We extracted influenza-related Twitter posts, which contained the words "influenza" or "flu" (corresponding words in Japanese, "インフルエンザ" and "インフル," were actually used). The period between November 2008 and April 2009 was defined as the "epidemic period of 2008–2009," that between November 2012 and June 2013 as the "epidemic period of 2012-2013," and that between April 2009 and July 2009 as the "pandemic period of 2009." Although there were actually two waves of the pandemic influenza in 2009 (spring and winter), we only used the first wave for our analyses because the second wave overlapped with the epidemic influenza season of 2009–2010.

2.1 Twitter post-based approach

The total number of influenza patients was calculated according to Twitter posts by influenza patients. Discriminating positive influenza posts from noise posts was conducted as a sentence classification task using natural language processing (NLP) similar to that used to filter spam e-mail. For this task, we implemented a straightforward influenza positive/noise discriminator for Twitter posts based on a machine learning approach, the Support Vector Machine (SVM) (Cortes and Vapnik, 1995). Sentences in each Twitter post were initially divided into words with a morphological analyzer JUMAN (`http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN`) to separate words. Six words both immediately before and after an influenza-related word in a Twitter post (12 words maximum) were selected as input for the SVM discriminator. The details of parameter tuning were described before (Aramaki et al., 2011). Training and performance evaluation was done through a 10-fold cross validation using 922 influenza-related Twitter posts in November 2009 (Twitter posts in this period were not used in the remainder of analyses in this paper), which were annotated manually by Japanese native speakers as either positive or noise influenza Twitter posts (caused 454 positive and 468 negative posts). The performance (F-measure, which is the harmonic mean of precision and recall) of this approach with ten-fold cross validation was 0.76. The number of positive Twitter posts divided by the total number of Twitter posts in Japanese in the same term was defined as the estimated relative number of influenza patients in Japan.

2.2 Google search query-based approach

The relative number of influenza patients estimated by Google was obtained from the Google Flu Trends (Japan Edition) website (`http://www.google.org/flutrends/jp/`). Ginsberg et al. described the algorithm (Ginsberg et al., 2009). In short, Google Flu Trends estimated the number of influenza patients based on the frequency of influenza-related Google search queries, which they found to have a high correlation with the number of patients who consulted physicians.

2.3 Observed number of influenza patients

We obtained the observed number of influenza patients in the epidemic periods of 2008-2009 and 2012-2013, and the pandemic period in 2009 in Japan from the official sentinel survey by the Infectious Disease Surveillance Center (IDSC; `http://idsc.nih.go.jp/`) at the National Institute of Infectious Disease (NIID) of Japan. The numbers of observed patients are based on weekly reports from around 5,000 fixed-point

medical facilities dispersed throughout Japan (about 2,000 internal medicine and 3,000 pediatric departments were selected) to the IDSC under the Law concerning the Prevention of Infections and Medical Care for Patients of Infections. Categorizing influenza-related Twitter posts, we manually classified 200 randomly selected influenza-related Twitter posts from the epidemic period of 2008-2009 and the pandemic period of 2009 into five categories: "Influenza positive" (Twitter posts from influenza patients), "Influenza negative" (includes negations, influenza-positive in the past but already recovered, and receiving vaccination), "Mention or joke about influenza," "News," and "Others."

3 Results

3.1 Performance of the social sensors for epidemic and pandemic influenza

During the seasonal epidemic periods of 2008–2009 and 2012-2013, the number of influenza patients estimated by both the Twitter post-based and the Google search query-based approaches showed good correlations (r=0.82 and r=0.93; r=0.82 and r=0.89) with the number of patients by the public surveillance system (Fig. 1).

However, during the pandemic period of 2009, both Twitter post-based and Google search query-based approaches showed extremely poor performance (r=-0.02 and r=0.23) as predictors of the number of influenza patients (Fig. 2). Although both approaches were able to react to the influenza pandemic, their responses were irrelevant. The Twitter post-based approach particularly over-responded to the pandemic.

3.2 Characterization of Twitter posts in the influenza epidemic and pandemic periods

To identify the cause of the failure in predicting the number of patients during the pandemic period, we sought to identify the differences between the properties of Twitter posts in the epidemic and pandemic periods.

The breakdown of 200 randomly selected Twitter posts that occurred during the two influenza outbreaks were examined (Fig. 3). The performances (F-measures) of the positive/noise discriminator against 200 randomly selected Twitter posts were 0.64 for the epidemic and 0.05 for the pandemic periods. During the pandemic period, the proportions of "Mention or joke about influenza," "News," and "Others" were much higher than those in the epidemic pe-

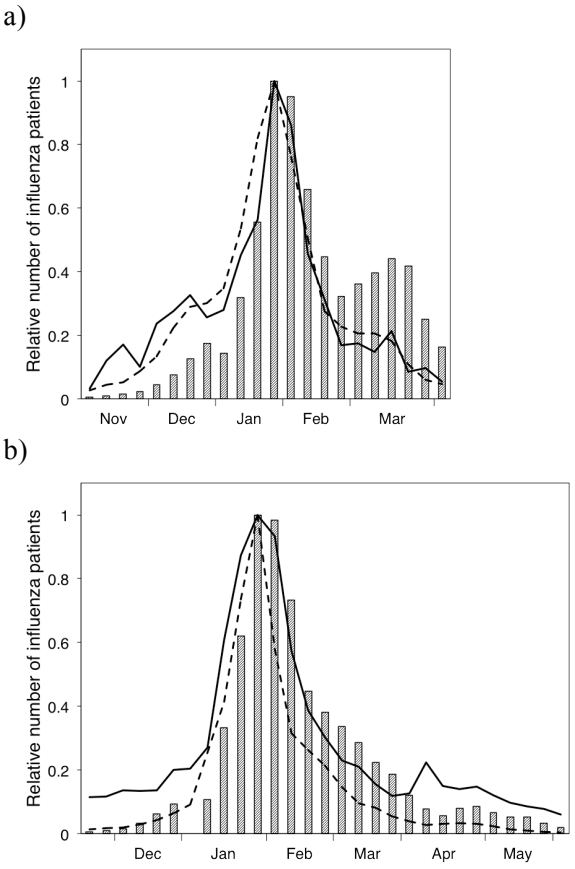


Figure 1. Trends in the relative number of influenza patients per week in the epidemic periods of a) 2008-2009 and b) 2012-2013 in Japan. The values were normalized with those at the peak of each season. The relative quantities of influenza patients per hospital are shown as bars. The estimated relative numbers by Twitter post-based approach are depicted as a solid line. Those by a Google search query-based approach are shown by the dotted line.

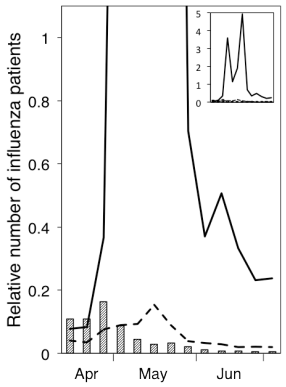


Figure 2. Trends in the relative number of influenza patients per week in the pandemic period of 2009 in Japan. The values were normalized with the same one in Fig. 1a. Data are presented in the same way as in Fig. 1.

riod, and 93% (185/200) of all randomly selected influenza-related Twitter posts fell into one of these three categories. Furthermore, the performances of the positive/noise discriminator were low for these three categories, especially for "Mention or joke about influenza."

4 Discussion

Both Twitter and Google are applicable to the early detection and survey of seasonal epidemic influenza because the correlation of the numbers of patients estimated by the Twitter post-based and Google search query-based approaches with those by the public surveillance system was high (Fig. 1). The Twitter-based and Google-based systems perform in real time, which can reduce the time lag of the current public surveillance systems, making it an actual feasible alternative that is one step ahead of the current official sentinel surveillance system. Because the numbers of Twitter and Google users are notably higher than the number of monitoring spots for the public surveillance system, Twitter and Google were able to monitor epidemics in higher geographic resolution, especially in densely populated areas, where infectious diseases can spread easily among people.

However, the Twitter post-based sensor caused a panic during the influenza pandemic. Moreover, even though it functioned better as a sensor than not responding at all to outbreaks, it completely failed to follow the trends of the outbreak (Fig. 2). If the performance of the positive/noise discriminator was sufficiently high, then such false-positive Twitter posts should have been removed properly. The actual performance was, however, very low in the pandemic period, with an F-measure of 0.05. A major cause of this catastrophic failure appears to be the types of Twitter posts observed frequently during this period. Of all influenza-related Twitter posts, 93% were classified as noise ("Mention and joke about influenza," "News," and "Others") (Fig. 3). The discriminator performance for these categories was poor, perhaps because the posts rarely contained frequently appearing expressions. It was difficult for a supervised machine learning approach when using bag-of-words as features to discriminate between positive data and noise.

Although the Google search query-based approach also got confused, it behaved much more moderately than the Twitter post-based approach did. An important difference between Twitter and Google is that Twitter is a kind of communication tool. Its users have the intention to have their posts read by other users. Studies in psychology have revealed that the following three conditions are related to rumor transmission: *personal anxiety*, *general uncertainty*, and *credulity* (trust in the rumor) (Rosnow, 1991). The environment was ripe for rumors to spread during the influenza pandemic period of 2009. A pandemic influenza in general is able to cause the death of millions of people. Also, it was initially reported that the mortality rate was extremely high in Mexico, and that the supply of vaccine was behind production schedule (*personal anxiety*). Infectability and mortality rates in developed countries were undetermined (*uncertainty*). The government continuously issued official announcements (*credulity*), though they were both unclear and several steps behind. Although some gap might exist in separating the number of Twitter posts and the rate of rumor transmission, the knowledge clarified in the traditional studies of social conversations is apparently applicable to studies of communication through Twitter. Microblogging is not the same as traditional social conversation, but it is apparently related to it. Therefore, we can assume that the amount of Twitter posts will explode under unusual situations such as life-threatening pandemics and bioterrorist attacks.

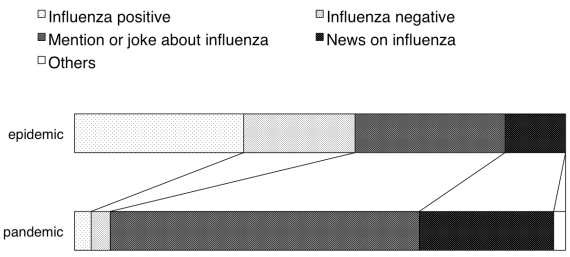


Figure 3. Comparison of types of influenza-related Twitter posts between the epidemic (upper) and the pandemic (lower) influenza periods. "Influenza positive" means Twitter posts from influenza patients. "Influenza negative" means Twitter posts about negation, positive in past but already recovered, and vaccination. The widths are proportional to the ratio of each type of post out of 200 randomly selected Twitter posts.

5 Conclusion

In this study, we have characterized two types of influenza sensors that were based on people's online behavior. Both the social media-based and

the search activity-based approaches could detect seasonal epidemic periods of influenza with fairly good performances, whereas the social media-based approach over-responded to the unexpected pandemic influenza. The number and type of posts on social media are likely to be affected by the condition for rumors to spread, which makes the social media-based approach less effective under such conditions.

Acknowledgements

The authors thank Kazutaka Baba for supporting software maintenance and Genta Kaneyama for providing us with Twitter data. This work was partially supported by the Precursory Research for Embryonic Science and Technology (PRESTO) program of the Japan Science and Technology Agency (JST).

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. EMNLP: 1568-1576.

Corinna Cortes and Vladimir N. Vapnik. 1995. Support-vector networks. Machine Learning 20: 273-297.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457: 1012-U1014.

Anette Hulth, Gustaf Rydevik, and Annika Linde. 2009. Web Queries as a Source for Syndromic Surveillance. PLoS ONE 4: e4378.

Heather A. Johnson, Michael M. Wagner, William R. Hogan, Wendy Chapman, Robert T. Olszewski, John Dowling, and Gary Barnas. 2004. Analysis of Web access logs for surveillance of influenza. Stud Health Technol Inform 107: 1202-1206.

Philip M. Polgreen, Yiling Chen, David M. Pennock, and Forrest D. Nelson. 2008. Using Internet Searches for Influenza Surveillance. Clinical Infectious Diseases 47: 1443-1448.

Ralph L. Rosnow. 1991. Inside rumor: A personal journey. American Psychologist 46: 484-496.

Alessio Signorini, Alberto M. Segre, and Philip M. Polgreen. 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. PLoS ONE 6: e19467.

Wilson Smith, C. H. Andrewes, and P. P. Laidlaw. 1933. A virus obtained from influenza patients. *Lancet* 2: 66-68.

World Health Organization. 2003. Influenza Fact sheet N°211. Available: http://www.who.int/mediacentre/factsheets/2003/fs211/en/. Accessed June 12, 2012.

World Health Organization. 2009. Influenza Fact sheet N°211. Available: http://www.who.int/mediacentre/factsheets/fs211/en/. Accessed June 12, 2012.

World Health Organization. 2010. What can I do? Available: http://www.who.int/csr/disease/swineflu/frequently_asked_questions/what/en/index.html. Accessed June 12, 2012.

T. Yamanouchi, K. Sakakami, and S. Iwashima. 1919. The infecting agent in influenza: An experimental research. *Lancet* 1: 971-971.

TogoStanza: Semantic Web framework for SPARQL-based data visualization in the biological context

Shinobu Okamoto
Database Center for Life Science
Research Organization of Information and systems, Japan
so@dbcls.rois.ac.jp

Shuichi Kawashima
Database Center for Life Science
Research Organization of Information and systems, Japan
kwsm@dbcls.rois.ac.jp

Takatomo Fujisawa
Center for Information Biology,
National Institute of Genetics
Research Organization of Information and Systems, Japan
tf@nig.ac.jp

Toshiaki Katayama
Database Center for Life Science
Research Organization of Information and systems, Japan
ktym@dbcls.jp

Over the last several decades, a huge amount of genomic and experimental information are aggregated as annotations on the genome sequences. Consequently, genome databases are still repeatedly designed and developed as a platform to integrate domain specific information. There are many overlapping generic features commonly required for genome databases. Therefore, the cost and time period of database development can be reduced by increasing reusability of the components that make up those genome databases. Moreover, to integrate the various biological databases and data sets, we have used semantic web technologies in this work.

TogoStanza (http://togogenome.org/stanza) is a simple framework supporting SPARQL queries and HTML rendering engine. It can be easily extended by JavaScript libraries to support various types of visualization patterns of biological data including table, bar chart, scattered plot, tree view, geographical map and genome browser. It provides an easy-to-use consistent framework for database developers and bioinformatics application programmers.

First, we developed RDF (Resource Description Framework) datasets to consolidate genomic information that vary in data types and data sources depending on the organisms (Katayama et al. 2013). Next, based on the RDF data, we developed 31 TogoStanza, reusable visualization components, optimized for each biological context including "environment", "organism" and "gene" (Figure 1). Finally, we also developed TogoGenome (http://togogenome.org/) which is a genome database and search interfaces as an application of TogoStanza. Then, we collaborated with MicorbeDB.jp (http://microbedb.jp/MDB) and CyanoBase (http://genome.microbedb.jp/cyanobase) database projects and each database already uses TogoStanza for visualizing data. The look and feel of each TogoStanza can be configured using CSS to enable its integration in different web applications.

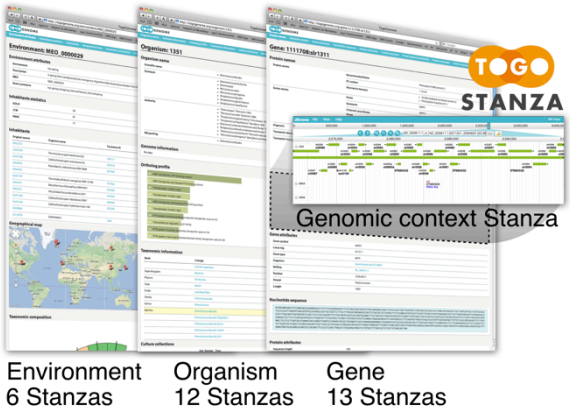


Figure 1. Illustration of the Web pages which are build up using the TogoStanza framework

References

Katayama T, et al. 2013. BioHackathon series in 2011 and 2012: penetration of ontology and Linked Data in life science domains. *J. Biomed Semantics* (Accepted)

Clinical Relation Extraction with Semi-Supervised Learning

Hiroki Ohba
Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku
Nagoya, Japan
sd12409@toyota-ti.ac.jp

Yutaka Sasaki
Toyota Technological Institute
2-12-1 Hisakata, Tempaku-ku
Nagoya, Japan
yutaka.sasaki@toyota-tia.c.jp

1 Introduction

In this study, we investigate methods to automatically extract medical entity relations from English clinical records. In our study, we take a semi-supervised approach. We used the Forth i2b2/VA (Informatics for Integrating Biology & the Bedside) shared task data [1]. To investigate the performance of relation extraction, we used correct concept annotations given in the dataset.

2 Classification method

Our target is relation between medical entity mentions for *medical problems*, *treatments*, and *tests*. The details of targeted relation categories are as follows:

- TrIP (*Treatment* improves *medical Problem*)
- TrWP (*Treatment* worsens *medical Problem*)
- TrCP (*Treatment* causes *medical Problem*)
- TrAP (*Treatment* is administered for *medical Problem*)
- TrNAP (*Treatment* is Not Administered because of *medical Problem*)
- PIP (*Medical Problem* indicates *medical Problem*)
- TeRP (*Test* Reveals *medical Problem*)
- TeCP (*Test* Conducted to investigate *medical Problem*)

We used SVM-light toolkit [3] for relation extraction. We choose the linear kernel because it has shown a good performance in many classification tasks. We use the lexical features, morphological features, and syntactic features to create input vectors. These features are selected in previous studies based on supervise learning. Since the SVM is a binary classification algorithm, we used the one-against-the-rest method for classification into the multiple relation categories.

3 Semi-Supervised Learning

There are several ways in employing semi-supervised learning. In this study, Self-Training [2] is employed. We first trained CRF-based medical named entity extraction and SVM-based relation extraction models on the i2b2 training data. Then, we applied these models to 62,269 unlabeled data in the i2b2 2010 dataset.

As we wanted to extract new reliable data from unlabeled data, we set a threshold. We experimented with a CRF-based NER's threshold between 0.90 and 0.99 by 0.01 step, and SVM-based relation extraction's threshold between -1.0 and 1.0 by 0.1 step. Then, we decided to set the threshold of 0.99 for CRF-based NER and 0 for SVM-based relation extraction, which generated 5,987 automatically labeled reliable data.

4 Experimental Results

We evaluated the results on the standard evaluation metrics: the recall, the precision, and the F-score. The result is shown in **Table1**.

Table1. Comparison of performance

	Recall	Precision	F-score
Supervised	0.6632	0.7440	0.7013
Semi-Supervised	0.6687	0.7416	0.7033

It is confirmed that the performance has been improved by employing Self-Training. Without external dictionary, the best i2b2 2010 relation extraction score was 0.6970. To further improve the performance, in the future work, we try to increase the size of unlabeled data.

Acknowledgement

This work was partly supported by JSPS KAKENHI Grant Number 25330271.

References

[1] Fourth i2b2/VA Shared-Task and Workshop: *Final Annotation Guidelines for Relations.* http://www.i2b2.org/NLP/Relations/Documentation.php .

[2] McClosky D, Charniak E, Johnson M. *Effective Self-Training for Parsing.* In: Proceedings of HLT/NAACL-2006, NY, USA, pp. 152-159, 2006.

[3] Thorsten Joachims, SVM-light Homepage http://svmlight.joachims.org/

A Unique Linear Representation of Carbohydrate Sequences for the Semantic Web

Issaku Yamada
The Noguchi Institute
Itabashi, Tokyo 173-0003
Japan
issaku@noguchi.or.jp

Kiyoko F. Aoki-Kinoshita
Department of Bioinformatics
Faculty of Engineering
Soka University
Hachioji, Tokyo 192-8557 Japan
kkiyoko@soka.ac.jp

1 Introduction

Due to the fast development of technologies for the Semantic Web in recent years, glycoinformatics developers of carbohydrate databases were faced with the problem of representing carbohydrate sequences (or glycans) uniquely in a linear format. In particular, glycans are complicated in that they are usually represented with ambiguity. That is, many times, if not most, all of the details are unknown or may be one of several linkages. Thus, the purpose of this work is to develop a unique linear representation of glycans such that they can be identified uniquely, whether or not they contain ambiguous information. We call this representation WURCS, for Web 3.0 Unique Representation of Carbohydrate Structures, and we describe version 1.0 here.

Carbohydrates take the form of branched, tree structures. In contrast to DNA, they cannot be sequenced because there is no template by which they are synthesized; they are synthesized by enzymes which add monosaccharides to the substrate one at a time. Consequently, they cannot be replicated, and so glycomics researchers who attempt to sequence glycans from biological samples are forced to work with small amounts. Because of this, glycan structures that are accumulated in glycan databases must often be represented ambiguously. To represent such structures on the Semantic Web, it is necessary to represent any such structure uniquely.

2 Method

We first focused on developing the framework for generating WURCS strings from a fully-defined glycan structure entered in atomic coordinates, such as a PDB (Berman et al., 2013) or MOL file (Dalby et al., 1992). Thus the main components of the glycan structures were first identified from the input: glycan, monosaccharides, modifications and aglycons. To define WURCS, the backbone, modifications and attachment sites were identified and each represented uniquely. Most other glycan formats, such as GlycoCT (Herget et al., 2008), use a dictionary to represent monosaccharides. However, we focused on defining monosaccharides which may not be commonly found in most glycan databases so that they can all be represented uniquely. To do so, we used what we call a SkeletonCode, similar to that used in MonosaccharideDB (http://www.monosaccharidedb.org) for each backbone to express them in a generic manner. We also defined what is called an ALIN, for Atomic LInear Notation to represent modifications, such as *N*-acetyl, sulfates, etc. We further defined COLINs, or COnnection LInear Notation, to represent the connection between the ALIN and SkeletonCode. Furthermore, we defined rules for each part to prioritize the components such that the generated string is ensured to be unique.

3 Results

We have developed the first unique linear notation for carbohydrate structures to theoretically cover all glycans that may be published in the literature. We have also formed an international working group to support the continual development of WURCS such that it can be adapted as an international standard, especially for its usage on the Semantic Web.

References

H.M. Berman, J.L. Markley, et al. 2013. *Biopolymers*, 99(3):218–222.

A. Dalby, J. Laufer, et al. 1992. *J. Chem. Inf. Comput. Sci.*, 32(3):244–255.

S. Herget, C.-W., von der Lieth, et al. 2008. *Carbohydr. Res.*, 343(12):2162-2171.

An Automatic Extractor for Biomedical Terms in Spanish

Leonardo Campillos-Llanos [*] **José Mª Guirao-Miras** [†] **Antonio Moreno-Sandoval** [*]
[*]Universidad Autónoma de Madrid, Madrid, Spain
[†]Universidad de Granada, Granada, Spain
{leonardo.campillos,antonio.msandoval}@uam.es
jmguirao@ugr.es

1 Introduction and background

Current non-commercial extractors for the medical domain are, for English, TerMine (Frantzi et al. 2000), or for Spanish, Vivaldi (2001). Here we present a hybrid system that combines lexically-based, tagger-based, and ruled-based methods. Our approach focuses on term classification.

2 System architecture

The system consists of four steps (Figure 1), each selecting different types of candidate terms:

- High reliability (single- and multi-word terms): we use a gold standard list of terms curated from medical dictionaries (e.g. Dorland 2005, RANM 2011).
- Medium reliability (single-word terms): we apply a silver standard list of terms that were not registered in dictionaries, but were found in medical books and articles. Those items that are not in the silver standard can be proposed as terms if: 1. a POS-tagger (GRAMPAL, Moreno and Guirao 2006) does not recognize them; and 2. a list of biomedical stems and affixes matches any unrecognized word.
- Medium reliability (multi-word terms): we use rules of multi-word term formation and phrase patterns.

In Moreno et al. (2013) we explain the methodology to collect the lists of terms from a corpus.

3 Conclusions and future work

Our tool provides an approach that is complementary to other extractors. However, domain experts have to further test and evaluate it. The system will be available at: http://cartago.lllf.uam.es/corpus3/index.pl.

Acknowledgements

The MultiMedica (TIN2010-20644-C03-03) and the MA2VICMR projects funded this work.

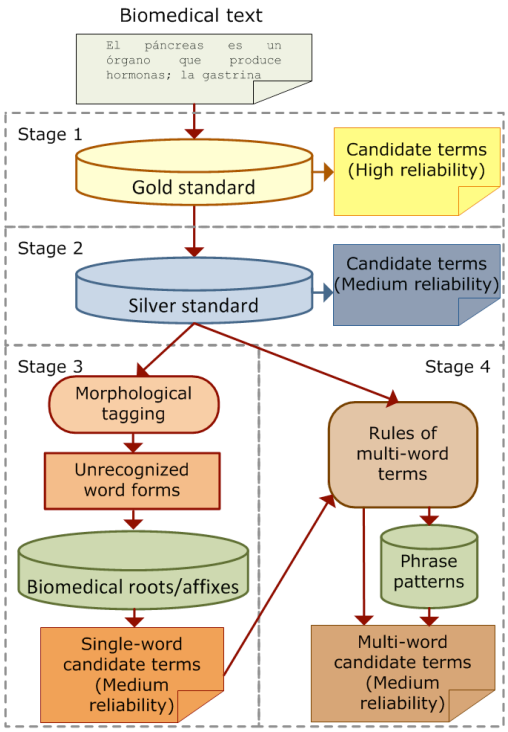


Figure 1. Processing pipeline

References

Dorland. 2005. *Diccionario enciclopédico ilustrado de medicina*. 30th edition. Madrid: Elsevier, D. L.

Frantzi, K., Ananiadou, S. and Mima, H. 2000. Automatic recognition of multi-word terms. *Intern. Journal of Digital Libraries*, 3(2): 117-132.

Moreno, A., and J. M. Guirao. 2006. Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus. In Y. Kawaguchi et al. (eds.) *Spoken Language Corpus and Linguistic Informatics*, 199-218. Amsterdam: John Benjamins.

Moreno, A., L. Campillos-Llanos, A. González, J. Mª. Guirao. 2013. An affix-based method for automatic term recognition from a medical corpus of Spanish. *Proc. 7th Corpus Linguistics 2013. Lancaster Univ.*

RANM (Royal National Academy of Medicine) 2011. *Diccionario de términos médicos*. Madrid: Editorial Médica Panamericana.

Vivaldi J. 2001. *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*. Univ. Politécnica de Catalunya.

OntoCloud – interactive visualization of relations between biomedical ontologies

Simon Kocbek
Database Center for Life Science, Research Organization of Information and Systems
Tokyo, Japan
simon@dbcls.rois.ac.jp

Jin-Dong Kim
Database Center for Life Science, Research Organization of Information and Systems
Tokyo, Japan
jdkim@dbcls.rois.ac.jp

Abstract

Ontologies in biomedicine are often used for standardization of biomedical terminology and can also be described as controlled biomedical vocabularies. Understanding structures of biomedical ontologies/vocabularies and their relations plays an important role in activities such as ontology reuse. In this paper we present OntoCloud, a tool for visualizing ontology relations and shapes.

1 OntoCloud

OntoCloud (http://bionlp.dbcls.jp/ontocloud/) is an interactive web application that offers a visualization of all biomedical ontologies available through BioPortal (Whetzel et al., 2011). The main goal of OntoCloud is to help ontology engineers and other users understand how BioPortal ontologies connect. OntoCloud groups ontologies that play an important role in connecting other ontologies and ontology communities. Users can identify which ontologies represent popular vocabularies that that are adopted for other vocabularies. In addition, OntoCloud aims to visualize which ontologies are closer to structured hierarchical vocabularies (i.e., contain more hierarchical relations like subclass_of, is_a or type_of), though illustrating ontology shapes. Users can interact with the application with the use of several functions, for example, searching for ontologies, visualize communities of densely connected ontologies, and visualize a custom subset of ontologies. Since BioPortal data often changes (e.g., new ontologies are being uploaded or new relations are being defined), OntoCloud also offers visualizations at different time points.

Figure 1 shows an example of visualizing a subset of ontologies in OntoCloud. Each node represents an ontology while edges represent connections between ontologies. Different groups of closely related ontologies are also represented with different colors. Structures of ontologies are represented with three different symbols (a circle, a triangle and a drop).

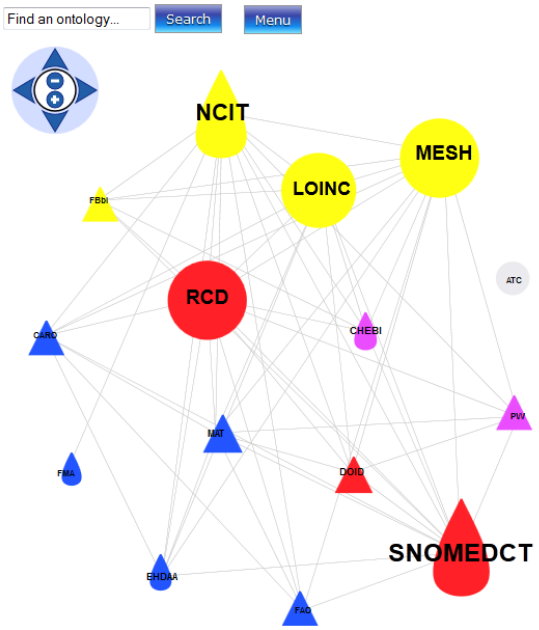


Figure 1: An example of visualization in OntoCloud.

Reference

Whetzel, P.L., Noy, N.F., Shah, N.H., et al.: BioPortal: enhanced function-ality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res;39:W541–5, (2011).

A New Approach of Extracting Biomedical Events Based on Double Classification

Xiaomei Wei[1,2], Kai Ren[1], Donghong Ji[1*]
[1]Wuhan University, Wuhan, China
[2]Huazhong Agriculture University, Wuhan, China
may@mail.hzau.edu.cn
dhji@whu.edu.cn

Abstract

Extraction of relations from literature is an important research topic in the field of biomedical natural language processing. Biomedical event is a kind of grained and complex relation. Recently, much research was focused on extracting biomedical events since it is proposed in BioNLP'09 challenges. Up to date, some approaches have been proposed. However, the performance of these approaches needs to be improved. In this paper we propose a novel method to extract biomedical events from text based on double classifications. In the first classification, we classify the candidate event pairs into 9 subsets in accord with 9 types of events. In the second classification, 9 binary classifiers are applied on each subset. Then it was evaluated on the Genia event extraction test datasets (Task 1) of BioNLP'2013, we get the performance with F-scores of 48.59 on the development dataset and 41.26 on the test dataset, respectively. In particular, we get high precision on all types of events while the recall is ordinary.

1 Method introduction

In our system, a complex biomedical event is defined as the combination of one or more ARG-TRIG pairs. Here ARG means the argument in the event and TRIG means the trigger word. In this article, we first extract the word string on the dependency path from the annotated protein to the root of syntax tree. Then we extract the ARG-TRIG pairs from the word string. In the following we obtain the positive ARG-TRIG pairs by double classifications which contain a multi-class classification and a set of binary classification. Finally a post-processing is applied to these pairs to construct events.

The workflow of the biomedical event extraction system is as follows:

(1) Text preprocessing
(2) Extracting candidate event pairs.
(3) Double classification to get the positive pairs.
(4) Post-processing to transfer the pairs into events.

The overall architecture of the system is shown in Figure. 1.

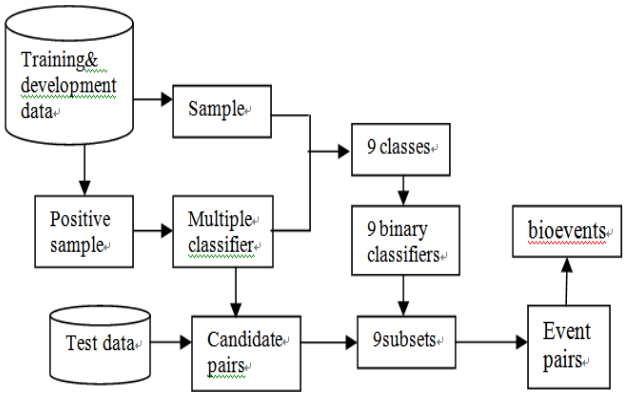


Figure .1 The overall architecture of the system

2 Experiment and evaluation

We use the datasets provided by BioNLP'2013 GE Shared Task to evaluate our extraction method. The datasets include training, development, and test data. The extracted events are submitted to the online evaluation system to evaluate the results.

Our system obtained F-score with 48.59 on the development dataset and 41.26 on the test dataset, respectively. We got the best precision which is comparable with all the systems participated in the BioNLP'13 challenge. In particular, the precision of Protein_catabolism reaches 100% while the recall is 50%. Obviously, compared with the f-score on development dataset, there is a dramatic decline on test dataset, which is due to our method of getting samples.

Acknowledgments

The work is funded by the following projects: Natural Science Foundation of China No.61133012 and No.61202304.

*Corresponding author

On Mention-level Gene Normalization

JoonYeob Kim **Seung-Cheol Baek** **Hee-Jin Lee** **Jong C. Park**
CS Department, KAIST, Korea
{jykim, scbaek, heejin, park}@nlp.kaist.ac.kr

Document-level gene normalization (DGN), which produces a list of gene identifiers relevant to an input document, helps database curators to search for articles of interest by indexing articles with gene identifiers. Recent advances in automatic extraction of information from the biology literature call for mention-level gene normalization (MGN) systems. However, there have been no annotated corpora for MGN, probably because of a somewhat unfounded assumption (*convertibility assumption*) that it might be straightforward to map gene mentions into gene identifiers given a list of gene identifiers for the document. In the present work, we constructed gold standard annotations for the MGN task and assessed the validity of the convertibility assumption with GeneTUKit (Huang et al., 2011), a state-of-the-art DGN system.

Since it is too costly to develop annotated corpora from scratch, we made the annotations on top of an existing corpus. We utilized the BioCreative2 GN corpus augmented by Hakenberg et al. (2008), which contains document-level annotations of genes of thirteen species. We first identified 1,139 gene mentions from 118 abstracts by using BANNER (Leaman and Gonzalez, 2008). We then paired each gene mention in an abstract with the gene identifiers annotated for the abstract, producing 3,939 mention-identifier pairs. Given a mention-identifier pair, human annotators first examined whether the gene mention is correctly identified. When the mention is correct, they checked also if the gene mention and the identifier represent the same gene to each other. Two human annotators majoring in bioinformatics achieved a sufficiently high inter-annotator agreement rate; they agreed upon 647 out of 796 mentions for the correctness of the gene mentions, and 1,774 out of 1,886 pairs for the appropriateness of mention-identifier pairing. The final version of annotations consists of 3,174 gene-mention pairs.

The resulting annotations show the following characteristics. First, contrary to our prediction that a mention is paired with at most one identifier, 15.4% of gene mentions are paired with two or more identifiers (e.g., homologous genes with different species), where 85% of them are paired with exactly two identifiers. This suggests that a single mention should be allowed to pair with at most two identifiers. Second, there are only 24 pairs of gene mentions with identical surface forms but with different gene identifiers, which are peculiar to the MGN task. This supports the convertibility assumption. Third, gene identifiers are paired with various numbers of gene mentions (3.7 ± 2.94), suggesting that there are relatively significant gene identifiers from the point of MGN systems.

Finally, we evaluated GeneTUKit for MGN with the help of two heuristic methods, or H1 and H2. H1 matches gene mentions with gene names produced by the GN system along with predicted gene identifiers, and H2 searches for the gene mentions from the known synonyms of predicted gene identifiers. The F-scores of MGN (0.377 with H1 and 0.337 with H2) are lower than the F-score for DGN (0.420). We will develop a MGN system by utilizing the characteristics identified earlier.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No.20110029447).

References

Hakenberg, J., Plake, C., Leaman, R., Schroeder, M., & Gonzalez, G. (2008). Inter-species normalization of gene mentions with GNAT. Bioinformatics, 24(16), i126-i132.

Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In Pacific Symposium on Biocomputing (Vol. 13, pp. 652-663).

Huang, M., Liu, J., & Zhu, X. (2011). GeneTUKit: a software for document-level gene normalization. Bioinformatics, 27(7), 1032-1033.

MPO: Microbial Phenotype Ontology for Comparative Genome Analysis

Shuichi Kawashima
Database Center for Life Science,
Research Organization of Information and systems, Japan
kwsm@dbcls.rois.ac.jp

Toshiaki Katayama
Database Center for Life Science,
Research Organization of Information and systems, Japan
ktym@dbcls.jp

Toshihisa Takagi
Department of Computational Biology,
the University of Tokyo, Japan
tt@k.u-tokyo.ac.jp

Shinobu Okamoto
Database Center for Life Science
Research Organization of Information and systems, Japan
so@dbcls.rois.ac.jp

In Database Center for Life Science (DBCLS), within the framework of the Life Science Database Integration Project, we are working on developing an infrastructure which may enable integrative use of life science databases. TogoGenome[1] is a data retrieval system developed within the project. Currently, various datasets related to microbial genome data are integrated in TogoGenome by using Semantic Web technology: all data are represented as RDF statements in TogoGenome. TogoGenome provides a faceted search system for microbial genes, allowing users to explore microbial genes interactively by applying multiple ontology terms as data filters.

In this study we focus on microbial phenotypes as new information to be introduced into TogoGenome. Phenotype is defined as the observable physical or biochemical characteristics of an organism resulting from both genetic makeup and environmental influences. Understanding how phenotypes emanate from a set of genotypes is one of the major goals of microbiology. Recently remarkable progress of technology has dropped the cost of genome sequencing drastically, which has led to an enormous amount of microbial genome data increasing at a rapid rate. On the other hand, information about microbial phenotypes has largely still been in the scientific papers or textbooks even though there has been a vast amount of descriptions over the past century. Therefore we started to develop MPO, an OWL ontology of microbial phenotype, and a set of phenotype LOD (Linked Open Data). These are integrated in TogoGenome to provide effective faceted search in the above regard.

Two data sources were used to collect terms relevant to microbial phenotypes. One is the Genome Online Database (GOLD). Each GOLD entry contains information related to phenotype in some fields (e.g. Oxygen requirement, Cell shape, Motility, Sporulation, Pressure, Temperature range, Salinity, Gram staining, Cell arrangement, Energy source, Metabolism and Phenotype). The other is a set of genome papers. Since biological features of the target species, which includes phenotypes, are often described in the introduction section of genome papers. We manually checked the phenotype terms described in the introduction section. Finally we defined 157 ontology terms from the collected terms.

Using a text-book and OMP, which is an another ontology of microbial phenotype developed in OBO format, as references, we defined a class which represents microbial phenotype at the first-level and seven classes at the second level in the hierarchy of MPO. The seven classes are Development, Environment condition tolerance, Microbial metabolism related phenotype, Morphology, Motility, Serotype and Staining. Then we assigned collected terms to the appropriate places in the hierarchy. MPO is available at BioPortal.

We also constructed a LOD data set that contains RDF statements of which the subjects are NCBI taxonomy ID and the objects are MPO terms. Currently we have been developing several web applications to visualize and analyze retrieved phenotype LOD, which are avaible at TogoGenome.

[1] http://togogenome.org/

Index of Authors