Third Workshop on Building and Evaluating Resources for Biomedical Text Mining

Workshop Programme

09:15 – 09:30 – Welcome (Sophia Ananiadou)

09:30 – 10:30 - Invited Talk (chair: Kevin Cohen)

Jun'ichi Tsujii, Microsoft Research Asia
Semantic and linguistic annotations in GENIA

10:30 – 11:00 - Coffee break

11:00 – 12:15 - Session 1 (chair: Paul Thompson)

11:00 – 11:25
Claudiu Mihăilă, Riza Theresa Batista-Navarro and Sophia Ananiadou
Analysing Entity Type Variation across Biomedical Subdomains

11:25 – 11:50
Suwisa Kaewphan, Sanna Kreula, Sofie Van Landeghem, Yves Van de Peer, Patrik R. Jones and Filip Ginter
Integrating Large-Scale Text Mining and Co-Expression Networks: Targeting NADP(H) Metabolism in E. coli with Event Extraction

11:50 – 12:15
Mariana Neves, Alexander Damaschun, Andreas Kurtz and Ulf Leser
Annotating and Evaluating Text for Stem Cell Research

12:15 – 14:00 Lunch break

14:00 – 15:15 - Session 2 (chair: Kevin Cohen)

14:00 – 14:25
Raheel Nawaz, Paul Thompson and Sophia Ananiadou
Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers

14:25 – 14:50
Fei Xia and Meliha Yetisgen-Yildiz
Clinical Corpus Annotation: Challenges and Strategies

14:50- 15:15
Dimitrios Kokkinakis
The Journal of the Swedish Medical Association - a Corpus Resource for Biomedical Text Mining in Swedish

15:15– 15:45 - Session 3 – Short poster presentations (chair: Paul Thompson)

15:15 – 15:25
Hercules Dalianis and Henrik Boström
Releasing a Swedish Clinical Corpus after Removing All Words - De-Identification Experiments with Conditional Random Fields and Random Forests

15:25- 15:35
Alyaa Alfalahi, Sara Brissman and Hercules Dalianis
Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus

15:35 – 15:45
Olfa Makkaoui, Julien Desclés and Jean-Pierre Desclés
Evaluation and Performance Improvement of the BioExcom System for the Automatic Detection of Speculation in Biomedical Texts

15:45 – 16:30 – Poster session and coffee break

16:30 – 17:20 - Session 4 (chair: Sophia Ananiadou)

16:30 – 16:55
Philippe Thomas, Tamara Bobić, Martin Hofmann-Apitius, Ulf Leser and Roman Klinger
Weakly Labeled Corpora as Silver Standard for Drug-Drug and Protein-Protein Interaction

16:55 – 17:20
Amber Stubbs
Developing Specifications for Light Annotation Tasks in the Biomedical Domain

17:20 – 17:30 - Concluding remarks (chair: Sophia Ananiadou)

Editors

Sophia Ananiadou	University of Manchester, UK
Kevin Cohen	University of Colorado School of Medicine, USA
Dina Demner-Fushman	National Library of Medicine, USA
Paul Thompson	University of Manchester, UK

Workshop Organizers

Sophia Ananiadou	University of Manchester, UK
Kevin Cohen	University of Colorado School of Medicine, USA
Dina Demner-Fushman	National Library of Medicine, USA
Paul Thompson	University of Manchester, UK

Workshop Programme Committee

Jari Björne	University of Turku, Finland
Olivier Bodenreider	National Library of Medicine, USA
Wendy Chapman	UCSD, USA
Hongfang Liu	Mayo Clinic, USA
Naoaki Okazaki	Tohoku University, Japan
Sampo Pyysalo	University of Manchester, UK
Andrey Rzhetsky	University of Chicago, UK
Stefan Schulz	Medical University Graz, Austria
Lucy Vanderwende	Microsoft Research, USA
Karin Verspoor	NICTA, Australia
John Wilbur	NCBI, NLM, NIH, USA
Stephen Wu	Mayo Clinic, USA
Pierre Zweigenbaum	LIMSI, France

Table of contents

Author Index

Introduction

This volume contains the papers accepted at the 3^{rd} *Workshop on Building and Evaluating Resources for Biomedical Text Mining* held at LREC 2012, Istanbul. Over the past decade, biomedical text mining has received a large amount of interest. Faced with the rapidly increasing volume of biomedical literature, domain experts have an ever-increasing need for tools that can help them locate isolate relevant nuggets of information from this deluge of information in a timely and efficient manner. The response to such issues by the natural language processing community can be clearly evidenced by the successful biomedical natural language processing workshops (BioNLP) that have been held over that past 10 years, in conjunction with ACL or NAACL meetings, to report the process in the field, as well as the founding of an ACL special interest group.

Biomedical text mining applications are reliant on high quality resources. These include databases and ontologies (e.g., Biothesaurus, UMLS Metathesaurus, MeSH and the Gene Ontology) and dictionaries/computational lexicons (e.g., the BioLexicon and the UMLS SPECIALIST lexicon). Recent years have also evidenced a large increase in the number of freely-available corpora (e.g., GENIA, GREC, AIMED, BioInfer, CRAFT, BioDRB) annotated with an expanding range of information types. These now include not only named entities and simple relations that hold between them, but also more complex event structures and coreference, as well as higher level information about how events are to be interpreted (e.g., facts, analyses, speculations, etc.) and discourse structure. Community shared tasks and challenges (e.g., JNLPBA, LL05, Biocreative I/II/III, BioNLP'09, BioNLP 2011, i2b2, etc.) also produce annotated corpora (on which the participating systems are trained and evaluated), in addition to steering research efforts to focus on open research problems. The development of high quality resources is very much relevant to META-NET (a Network of Excellence consisting of 54 research centres from 33 countries), that aims to stimulate a pan-European acceleration of research language technologies; this is dependent on the availability of appropriate resources.

The papers in this volume exemplify the diversity of research that is currently taking place. Three papers concern resources for a relatively resource-poor language, i.e. Swedish. One of these describes a biomedical corpus derived from the Journal of the Swedish Medical Association (Kokkinakis), whilst the other two address de-identification of records in Swedish clinical corpora to remove Protected Health Information (PHI), using 2 different methods, i.e. pseudonymysation (Alfalahi et al.) and replacement of words with features (Dalianis and Boström). A third paper considering clinical corpora (Xia and Yetisgen-Yildiz) explains the challenges faced during annotation, and highlights the need for domain experts and detailed guidelines. In contrast, a further paper about annotation (Stubbs) proposes an annotation methodology for "light" annotation tasks for biomedical corpora, which do not require extensive training or exceptionally long annotation periods.

Three papers relate to biomedical relations or events. Kaewphan et al. describe the application of a literature-scale event extraction resource, EVEX, to NADP(H) metabolism regulation in *Escherichia coli*. The other two papers present new annotated corpora. Thomas et al. present two new corpora for protein-protein interactions and drug-drug interactions, which were automatically annotated, using distant supervision methods. Nawaz et al. describe the application of their multi-dimensional meta-knowledge annotation scheme to previously annotated biomedical events in a small collection of full papers, in order to enrich them with aspects of event interpretation such as negation, speculation, and knowledge source. The results are compared with a previous annotation effort for abstracts. The importance of recognising such interpretative information in biomedical texts is reinforced in the paper by Makkaoui et al., which evaluates a system for annotating speculative sentences on the BioScope corpus.

The remaining two papers in this volume concern named entity annotations. Neves et al. present a corpus for stem cell research, which is annotated with different types of entities relevant to this subdomain. Preliminary results of automatic recognition of these entities are also presented.

Mihăilă et al. examine the distribution of named entity types across 20 different biomedical subdomains. The degree of difference or similarity between different subdomains can be an important consideration when adapting automated tools from one subdomain to another.

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

Sophia Ananiadou, Kevin Cohen, Dina Demner-Fushman and Paul Thompson

Analysing Entity Type Variation across Biomedical Subdomains

Claudiu Mihăilă, Riza Theresa Batista-Navarro, Sophia Ananiadou

National Centre for Text Mining
School of Computer Science, University of Manchester
Manchester Interdisciplinary Biocentre,
131 Princess Street, M1 7DN, Manchester, UK
Email: {claudiu.mihaila, riza.batista-navarro}@cs.man.ac.uk,
sophia.ananiadou@manchester.ac.uk

Abstract

Previous studies have shown that various biomedical subdomains have lexical, syntactic, semantic and discourse structure variations. It is essential to recognise such differences to understand that biomedical natural language processing tools, such as named entity recognisers, that work well on some subdomains may not work as well on others. In this paper, we investigate the pairwise similarity (or dissimilarity) amongst twenty selected biomedical subdomains, at the level of named entity types. We evaluate the contribution of these types in the classification task by computing the chi-squared statistic over their distributions. We then build a binary classifier for each possible pair of subdomains, the results of which indicate the subdomains that are highly different or similar to others. The findings can be of potential use to those building or using named entity recognisers in determining which types of named entities need to be taken into consideration or in adapting already existing tools.

Keywords: named entity, subdomain variation, machine learning, biomedical text mining

1. Introduction

Statements regarding associations and connections between biological events and processes are central to identifying facts and claims of interest in biomedical science. Both events and processes are created on top of biological entities, so it is necessary to recognise the latter with the highest possible precision. Thus, the development of tools and resources for the automatic analysis of named entities (NEs) is key to information extraction (IE) and text mining for domain-specific scientific text.

In the past decade, researchers have focussed on fundamental tasks needed to create intelligent systems capable of improving search engine results and easing the work of biologists. More specifically, researchers have concentrated mainly on named entity recognition, normalisation to specialised databases (Krallinger et al., 2008) and extracting simple binary relations between entities.

Whilst a multitude of tools and resources have been introduced in domain-specific natural language processing (NLP) efforts for the recognition of entity mentions in text, a high proportion of these was trained and evaluated on popular corpora such as BioInfer (Pyysalo et al., 2007), GENETAG (Tanabe et al., 2005), GENIA (Kim et al., 2008), and PennBioIE (Kulick et al., 2004), as well as shared task corpora from BioCreative I, II, III (Arighi et al., 2011) and BioNLP 2009 and 2011 (Kim et al., 2011). Most of these corpora consist of documents from the molecular biology subdomain. However, previous studies (discussed in Section 2) have established that different biomedical sublanguages exhibit linguistic variations. It follows that tools which were developed and evaluated on corpora derived from one subdomain might not always perform as well on corpora from another subdomain. Understanding these linguistic variations is essential to domain adaptation of natural language processing tools.

In this paper, we highlight the similarities and differences found between biomedical sublanguages by focussing on the various types of named entities that are relevant to them. We show that for some pairs of subdomains, the frequencies of their named entity types are very similar, implying that these subdomains are very closely related. For others, however, the frequencies of different named entity types are diverse enough to allow a classifier for biomedical subdomains to be built based upon them.

This study is performed on open access journal articles found in the UK PubMed Central (UKPMC) (McEntyre et al., 2010), an article database that extends the functionality of the original PubMed Central (PMC) repository[1]. This database was chosen as our source, as most of the documents it contains are already tagged with named entity information. Reported in this paper are results obtained for 8,000 articles from 20 different biomedical subdomains.

2. Related Work

The work of Harris (1968) introduced a formalisation of the notion of sublanguage, which he defined as a subset of general language. According to his theory, it is possible to process specialised languages, since they have a structure that can be expressed in a computable form. Several works on the study of biomedical languages substantiated his theory, including the work of Sager et al. (1987) on pharmacological literature and lipid metabolism, and that of Friedman et al. (2002) analysing the properties of clinical and biomolecular sublanguages.

[1] http://www.ncbi.nlm.nih.gov/pmc

Other studies have investigated the differences between general and biomedical languages by focussing on specific linguistic aspects, such as verb-argument relations and pronominal anaphora. For instance, Wattarujeekrit et al. (2004) analysed the predicate-argument structures of 30 verbs used in biomedical articles. Their results suggest that, in certain cases, a significant difference exists in the predicate frames compared to those obtained from analysing news articles in the PropBank project (Palmer et al., 2005). Similarly, based on the GENIA and PennBioIE corpora, Cohen et al. (2008) perform a study of argument realisation with respect to the nominalisation and alternation of biomedical verbs. They conclude that there is a high occurrence of these phenomena in this semantically restricted domain, and underline that this sublanguage model applies only to biomedical language.

Taking a different angle, Stetson et al. (2002) uncovered the differences between "signout" notes and other medical notes (e.g., ambulatory clinic notes and discharge summaries) in terms of three aspects: discourse length, abbreviation use and abbreviation ambiguity. Based on their findings, "signout" notes are shorter and use a higher number of less ambiguous abbreviations. Nguyen and Kim (2008), on the other hand, examined the differences in the use of pronouns in general and biomedical domains by studying the MUC, ACE and GENIA corpora. They observed that compared to the MUC and ACE corpora, the GENIA corpus has significantly more occurrences of neutral and third-person pronouns, whilst first and second person pronouns are non-existent.

Verspoor et al. (2009) measured the lexical and structural variation in biomedical Open Access journals and subscription-based journals, concluding that there are no significant differences between them. Therefore, a model trained on one of these sources can be used successfully on the other, as long as the subject is maintained. Furthermore, they compare a mouse genomics corpus with two reference corpora, one composed of newswire texts and another of general biomedical articles. In this case, unsurprisingly, significant differences are found across many linguistic dimensions. Relevant to our study is the comparison between the more specific mouse genome corpus to the more general biomedical one: whilst similar from some points of view, such as negation and passivisation, they differ in sentence length and semantic features, such as the presence of various named entities.

This study, in contrast, investigates the differences and similarities between any two of twenty biomedical sublanguages at the level of named entities. Examining the distributions of different named entity types across several categories, our work is subtly similar to that of Cohen et al. (2010) who looked at the distributional variations of semantic classes in their effort to characterise the differences between abstracts and full texts. Four semantic classes, namely, *Gene*, *Mutation*, *Drug* and *Disease*, were taken into account in their study. Except for *Gene*, significant differences in terms of densities per thousand words have been observed between abstracts and full texts.

Also relevant is the work of Lippincott et al. (2011) in which a clustering-based quantitative analysis of the linguistic variations across 38 different biomedical sublanguages was presented. They investigate four dimensions relevant to the performance of NLP systems, i.e. vocabulary, syntax, semantics and discourse structure. With regard to semantic features, the authors induced a topic model using Latent Dirichlet Analysis for each word, and then extended the model to documents and subdomains according to observed distributions. Their conclusion is that an unsupervised machine learning system is able to create robust clusters of subdomains, thus proving their hypothesis that the commonly used molecular biology subdomain is not representative of the domain as a whole. In contrast, we examine the differences and similarities between biomedical sublanguages at the level of named entities, using supervised machine learning algorithms and on a different number of subdomains.

3. Methodology

We initially created a corpus of documents from various biomedical subdomains, from which we then extracted named entity information automatically. The NEs were later transformed into input for machine learning algorithms, as discussed below.

3.1. Document Collection

A corpus was created by first searching the NLM Catalog[2] for journals which are in English and available via PubMed Central, and then narrowing down the results to those whose Broad Subject Term attributes contain only one biomedical subdomain name. Since we are interested in full-text articles, we retained only those journals which are available within the PubMed Open Access subset[3]. After obtaining the total number of documents across different journals in each subdomain, we retained only those subdomains with at least 400 documents.

Using the PMC IDs of all articles under the 20 remaining subdomains, we retrieved documents from UKPMC. For each subdomain, we selected the first 400 documents with the largest number of annotated named entities. The retrieved documents are in XML format. Several unusable fragments were removed before converting them to plain text. Examples of such fragments are article metadata (authors, affiliations, publishing history), tables, figures, and references. Table 1 shows the 20 subdomains and the approximate size of the corresponding corpus subset (in number of words) after the pre-processing step.

3.2. Tagging of Named Entities

We formed a silver standard corpus by harmonising the annotations of multiple resources and named entity recognisers. This method was chosen due to the fact that there are no gold standard annotations available for such a large number of full-text articles.

To create the named-entity-tagged corpus, we used a simple method that augments the named entities present in the UKPMC articles with the output of two named entity recognition tools (NERs), i.e. NeMine and OSCAR. In UKPMC,

[2] http://www.ncbi.nlm.nih.gov/nlmcatalog

[3] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist

Subdomain	Shortname	No. of words
Allergy and Immunology	Allergy	0.9M
Biology	Biology	3.3M
Cell Biology	CellBio	3.2M
Communicable Diseases	Communi	1.4M
Critical Care	Critica	1.6M
Environmental Health	Environ	1.9M
Genetics	Genetic	3.0M
Health Services Research	HealthS	1.7M
Medical Informatics	Medical	2.6M
Medicine	Medicin	2.1M
Microbiology	Microbi	2.6M
Neoplasms	Neoplas	2.2M
Neurology	Neurolo	2.3M
Pharmacology	Pharmac	1.8M
Physiology	Physiol	3.5M
Public Health	PublicH	1.7M
Pulmonary Medicine	Pulmona	1.9M
Rheumatology	Rheumat	1.9M
Tropical Medicine	Tropica	1.7M
Virology	Virolog	2.3M

Table 1: The 20 subdomains in the corpus, their shortnames and number of words in the corpus subset.

only six named entity types are annotated; with the use of NeMine and OSCAR, however, we obtained a total of 19 different classes of entities, summarised in Table 2.

Named entities in the UKPMC database were identified using NeMine (Sasaki et al., 2008), a dictionary-based statistical named entity recognition system. This system was later extended and used by Nobata et al. (2009) to include more types, such as phenomena, processes, organs and symptoms. We used this most recent version of the software as our second source of more diverse entity types.

The Open-Source Chemistry Analysis Routines (OSCAR) software (Corbett and Copestake, 2008; Jessop et al., 2011) is a toolkit for the recognition of named entities and data in chemistry publications. Currently in its fourth version, it uses three types of chemical entity recognisers, namely regular expressions, patterns and Maximum Entropy Markov models.

Nevertheless, due to the combination of several NERs, some NE types are more general and comprise other more specific types, therefore leading to double annotation. For instance, the *Gene|Protein* type is more general than both *Gene* and *Protein*, so only *Gene* or *Protein* will be kept in case they overlap with *Gene|Protein*. The same applies to the *Chemical molecule* type, which is a hypernym of *Gene*, *Protein*, *Drug* and *Metabolite*. In the case of multiple annotations over the same span of text, we removed the more general *Chemical molecule* type, so that each entity is labelled only with the more specific category assigned. Although this type of multiple annotations was frequent, we did not encounter any case of contradicting annotations over the same span of text.

This corpus is available upon request from the authors.

Type	UKPMC	NeMine	OSCAR
Gene	✓	✓	
Protein	✓	✓	
Gene|Protein	✓		
Disease	✓	✓	
Drug	✓	✓	
Metabolite	✓	✓	
Bacteria		✓	
Diagnostic process		✓	
General phenomenon		✓	
Indicator		✓	
Natural phenomenon		✓	
Organ		✓	
Pathologic function		✓	
Symptom		✓	
Therapeutic process		✓	
Chemical molecule			✓
Chemical adjective			✓
Enzyme			✓
Reaction			✓

Table 2: Named entity types and their source.

3.3. Experimental Setup

Based on the corpus previously described, we created a data set for supervised machine learning algorithms. Every document in the corpus was transformed into a vector consisting of 19 features. Each of these features corresponds to an entity type in Table 2, having a numeric value ranging from 0 to 1. This value represents the ratio of the specific entity type to the total number of named entities recognised in that document, as shown in Equation 1.

$$\theta = \frac{n_{type}}{N} \tag{1}$$

, where n_{type} represents the number of named entites of a certain type in a document and N represents the total number of named entities in that document. Each vector was labelled with the name of the subdomain to which the respective document belongs.

From the twenty subdomains in the corpus, we formed all possible combinations of two (thus resulting in a total of 190 pairs) for each of which we built a binary classifier. Weka (Witten and Frank, 2005; Hall et al., 2009) was employed as the machine learning framework, due to its large variety of classification algorithms. We experimented with a large number of classifiers, including J48, JRip, Logistic, RandomTree, RandomForest, SMO and combinations of these with AdaBoost. Evaluation was performed using the 10-fold cross-validation technique. RandomForest obtained the best F-score in 86 out of the 190 subdomain pairs, whilst the best result in 98 cases was obtained by AdaBoost in combination with other algorithms (JRip, RandomTree, Logistic). The remaining pairs were best classified by JRip (4 pairs) and Logistic (2 pairs). We therefore decided to present in this paper only the results using RandomForest.

4. Results and Analysis

We initially evaluated the value of the selected features for our task with a statistical significance test, and then performed the machine-learning experiments. Finally, we discuss the obtained results.

4.1. Feature Evaluation

To confirm the value of the selected features in classifying documents into subdomains, we performed the chi-squared (χ^2) test of independence between each named entity and each pair of subdomains. Chi-squared is defined in Equation 2, whilst the expected value of the observation is computed according to Equation 3.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \qquad (2)$$

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,k} \sum_{k=1}^{r} O_{k,j}}{N} \qquad (3)$$

The values are obtained by applying the ChiSquare Attribute Evaluator that is implemented in Weka. Each result contains a vector of 19 chi-squared scores, one for each feature. To visualise this graphically, we computed the Frobenius norm of the vector of chi-squared values for each subdomain pair. The Frobenius norm is defined as the square root of the sum of the absolute squares of its elements, as seen in Equation 4 (Golub and van Van Loan, 1996).

$$\|A\|_F = \sqrt{AA^*} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \qquad (4)$$

, where A^* denotes the conjugate transpose of A.
The resulting heatmap is included as Figure 1. The higher the value of the Frobenius norm, the better is the combination of features for distinguishing between the two subdomains in the pair.
To gain an insight into which features contribute most or least to the overall task, the sum of the chi-squared statistic for each feature was taken over all pairs of subdomains. We present the maximum and minimum values obtained from this exercise in Table 3.

4.2. Classifier Results

From the 20 subdomains, a binary classifier was built for each possible subdomain pair, as discussed in the previous section. The heatmap in Figure 2 shows the performance of each of the 190 pairs in terms of F-score. This heatmap is non-symmetric, in the sense that the F-score of subdomains A and B is different from that of B and A. All F-scores presented in this heatmap are computed with respect to the subdomain on the Y-axis (left) and against the subdomains on the X-axis (top).
A cell with a dark shade of grey corresponds to a pair of subdomains which are discernible from each other by a classifier trained on named entity type frequencies. *Cell Biology* and *Pharmacology*, for example, are found to have very distinct named entity type frequencies, as evidenced by the very good performance (97.15% F-score) of the classifier for them.

Type	Mean
Bacteria	10.57
Chemical adjective	19.07
Chemical molecule	87.84
Diagnostic process	24.30
Disease	195.06
Drug	82.57
Enzyme	30.77
Gene	78.03
Gene|Protein	145.94
General phenomenon	0.34
Indicator	63.10
Metabolite	112.17
Natural phenomenon	7.07
Organ	35.78
Pathologic function	5.79
Protein	140.83
Reaction	108.43
Symptom	16.46
Therapeutic	56.09

Table 3: Mean values of the chi-squared statistic for each feature over all pairs of subdomains.

On the other hand, a lighter tint of grey means that the corresponding pair consists of subdomains which are very similar in their named entity type frequencies. Such is true in the case of *Communicable Diseases* and *Tropical Diseases*, for instance, in which the classifier obtained an F-score of 56.63%.

4.3. Analysis

From these results, we are able to enumerate the subdomains which can be considered as different or similar to a subdomain of interest in terms of frequencies of their named entity types. In obtaining the most similar subdomains, we looked at the pairs whose F-score is at the lower end of the scale. There are no pairs for which the F-scores are between 50 to 55%, and only two pairs fall within the 55-60%-range. We hence used as threshold an F-score of 65% (i.e., subdomains in pairs for which the F-score of the classifier is 65% and below were considered similar). On the other hand, we looked at the other end of the scale (i.e., pairs for which the F-score of the classifier is 95% and above) to obtain a listing of the most dissimilar subdomains.
Findings in Table 4 suggest that when building NLP tools (e.g., named entity recognisers) for documents under the subdomain in the first column, one might trivially adapt those developed for the corresponding subdomains in the second column. A named entity recogniser for the *Microbiology* subdomain, for example, might be trivially applied to *Neoplasms* documents. However, it might also be the case that there are no named entity recognisers built yet that are specialised for these subdomains.
In contrast, those built for the subdomains in the second column of Table 5 might need further training or adaptation in applying them to the corresponding subdomain in the first column, as these tools might have been trained on

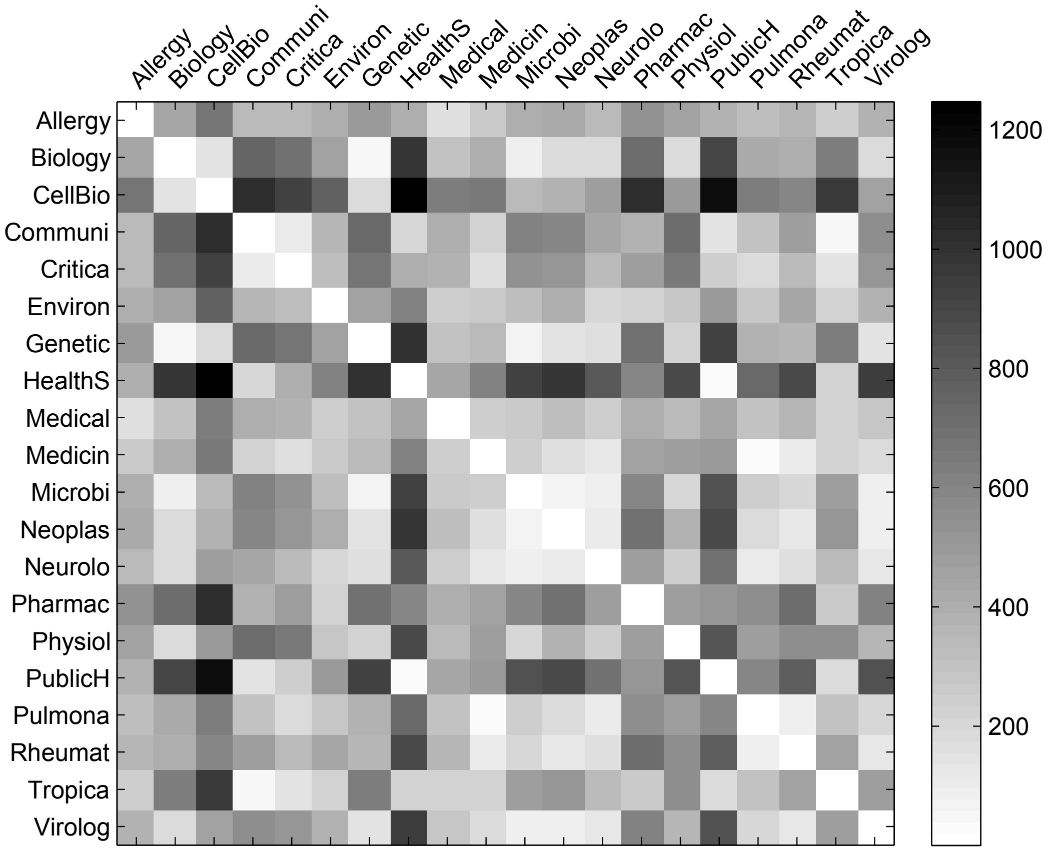


Figure 1: A heatmap showing the Frobenius norm based on the chi-squared vector for each pair of subdomains.

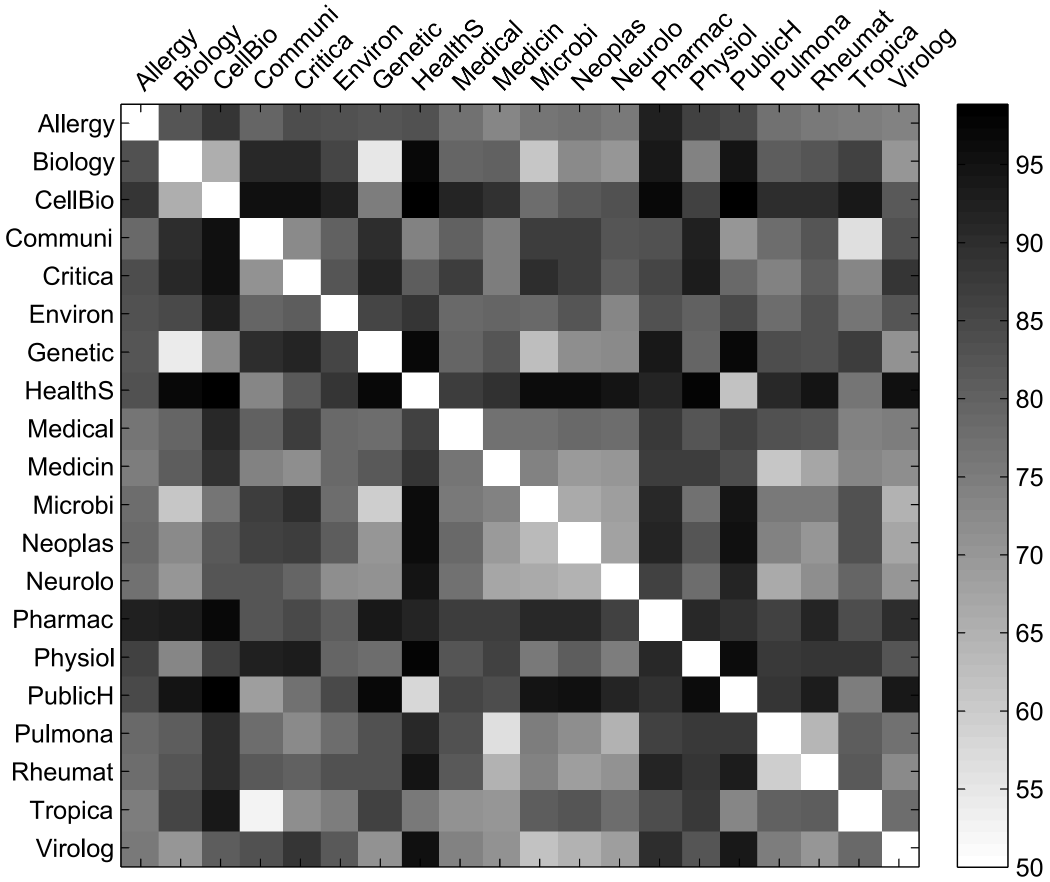


Figure 2: A heatmap showing the performance (in F-score) of each classifier built for each pair of subdomains.

Subdomain	Similar subdomains
Biology	Cell Biology, Genetics, Microbiology
Communicable Diseases	Tropical Diseases
Medicine	Pulmonary Medicine
Health Services Research	Public Health
Genetics	Microbiology
Pulmonary Medicine	Rheumatology
Microbiology	Virology

Table 4: Similar subdomains. The subdomains listed in the second column can be considered as highly similar to the corresponding subdomain in the first column based on their named entity type frequencies.

Subdomain	Dissimilar subdomains
Biology	Public Health, Health Services Research
Cell Biology	Critical Care, Communicable Diseases, Pharmacology, Public Health, Health Services Research
Genetics	Public Health, Health Services Research
Health Services Research	Microbiology, Neoplasms, Physiology, Rheumatology, Virology
Neoplasms	Public Health
Physiology	Public Health

Table 5: Dissimilar subdomains. The subdomains listed in the second column can be considered as different from the corresponding subdomain in the first column based on their named entity type frequencies.

documents where the named entity types which occur frequently in the subdomain of interest, are sparse. For instance, there is no certainty that NERs developed for the *Pharmacology* domain will work well on *Neoplasms* documents.

We computed the mean along each row and column of the heatmap, and determined that both the row and column corresponding to *Medicine* produced the minimum, while *Pharmacology* has the maximum. This finding suggests that *Medicine* is the biomedical subdomain which is most “alike” every other subdomain, irrespective of the direction F-score is computed in, while *Pharmacology* is the least one. In developing a named entity recogniser for *Pharmacology*, one has to consider its differences with other biomedical subdomains in terms of named entity type distributions.

5. Conclusion

We formed a silver standard corpus from 20 biomedical subdomains and built a binary classifier for each possible subdomain pair. From the results, we have observed which subdomains are highly discernible from each other by a classifier, in terms of named entity type frequencies. However, there are also cases when a classifier is unable to distinguish between subdomains, implying that they have highly similar named entity type distributions.

Such differences and similarities in named entity type frequencies should be considered when developing automated tools for one subdomain and adapting them for use on another.

6. References

Cecilia Arighi, Zhiyong Lu, Martin Krallinger, Kevin Cohen, W Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy Wu. 2011. Overview of the BioCreative III Workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Kevin Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. 2008. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9):e3158, 09.

Kevin Bretonnel Cohen, Helen Johnson, Karin Verspoor, Christophe Roeder, and Lawrence Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

Peter Corbett and Ann Copestake. 2008. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics*, 9(Suppl 11):S4.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Gene H. Golub and Charles F. van Van Loan. 1996. *Matrix Computations (Johns Hopkins Studies in Mathematical Sciences) (3rd Edition)*. The Johns Hopkins University Press, 3rd edition, October.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

Zellig Harris. 1968. *Mathematical Structures of Language*. John Wiley and Son, New York.

David Jessop, Sam Adams, Egon Willighagen, Lezan Hawizy, and Peter Murray-Rust. 2011. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(1):41.

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

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA, June. Association for Computational Linguistics.

Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. 2008. Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biology*, 9(Suppl 2):S1.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, and Lyle Ungar. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of the BioLINK 2004*.

Thomas Lippincott, Diarmuid Seaghdha, and Anna Korhonen. 2011. Exploring subdomain variation in biomedical language. *BMC Bioinformatics*, 12(1):212.

Johanna R. McEntyre, Sophia Ananiadou, Stephen Andrews, William J. Black, Richard Boulderstone, Paula Buttery, David Chaplin, Sandeepreddy Chevuru, Norman Cobley, Lee-Ann Coleman, Paul Davey, Bharti Gupta, Lesley Haji-Gholam, Craig Hawkins, Alan Horne, Simon J. Hubbard, Jee-Hyub Kim, Ian Lewin, Vic Lyte, Ross MacIntyre, Sami Mansoor, Linda Mason, John McNaught, Elizabeth Newbold, Chikashi Nobata, Ernest Ong, Sharmila Pillai, Dietrich Rebholz-Schuhmann, Heather Rosie, Rob Rowbotham, C. J. Rupp, Peter Stoehr, and Philip Vaughan. 2010. UKPMC: a full text article resource for the life sciences. *Nucleic Acids Research*.

Ngan L. T. Nguyen and Jin-Dong Kim. 2008. Exploring domain differences for the design of pronoun resolution systems for biomedical text. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 625–632, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chikashi Nobata, Yutaka Sasaki, Naoaki Okazaki, C.J. Rupp, Jun'ichi Tsujii, and Sophia Ananiadou. 2009. Semantic search on digital document repositories based on text mining results. In *International Conferences on Digital Libraries and the Semantic Web 2009 (ICSD2009)*, pages 34–48.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Bjorne, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2007. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1):50.

Naomi Sager, Carol Friedman, and Margaret Lyman. 1987. *Medical Language Processing: Computer Management of Narrative Data*. Addison-Wesley, Reading, MA.

Yutaka Sasaki, Yoshimasa Tsuruoka, John McNaught, and Sophia Ananiadou. 2008. How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics*, 9(Suppl 11):S5.

Peter D Stetson, Stephen B Johnson, Matthew Scotch, and George Hripcsak. 2002. The sublanguage of cross-coverage. *Proceedings of the AMIA Symposium*, pages 742–746.

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

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

Tuangthong Wattarujeekrit, Parantu Shah, and Nigel Collier. 2004. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(1):155.

Ian Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann.

Integrating Large-Scale Text Mining and Co-Expression Networks: Targeting NADP(H) Metabolism in E. coli with Event Extraction

Suwisa Kaewphan[*], **Sanna Kreula**[†], **Sofie Van Landeghem**[‡],
Yves Van de Peer[‡], **Patrik R. Jones**[†], **Filip Ginter**[*]

[*]Department of Information Technology, University of Turku
Joukahaisenkatu 3-5B, 20520 Turku, Finland
sukaew,figint@utu.fi

[†]Bioenergy group, University of Turku
Tykistökatu 6A, 6krs, 20520 Turku, Finland
sanmpe,patjon@utu.fi

[‡]Department of Plant Systems Biology, VIB,
Department of Plant Biotechnology and Bioinformatics, Ghent University
Technologiepark 927, 9052 Gent, Belgium
yves.vandepeer,sofie.vanlandeghem@psb.vib-ugent.be

Abstract

We present an application of EVEX, a literature-scale event extraction resource, in the concrete biological use case of NADP(H) metabolism regulation in *Escherichia coli*. We make extensive use of the EVEX event generalization based on gene family definitions in Ensembl Genomes, to extract cross-species candidate regulators. We manually evaluate the resulting network so as to only preserve correct events and facilitate its integration with microarray-based co-expression data. When analysing the combined network obtained from text mining and co-expression, we identify 41 candidate genes involved in triangular patterns involving both subnetworks. Several of these candidates are of particular interest, and we discuss their biological relevance further. This study is the first to present a real-world evaluation of the EVEX resource in particular and literature-scale application of the systems emerging from the BioNLP Shared Task series in general. We summarize the lessons learned from this use case in order to focus future development of EVEX and similar literature-scale resources.

Keywords: event extraction, EVEX, NADP(H), co-expression

1. Introduction

The field of natural language processing in the biomedical domain (BioNLP) aims at supporting life science research in dealing with the mass of available scientific literature in an efficient manner. Typical use cases for BioNLP include support for biological database curation, efficient retrieval of articles relevant to a particular biomedical molecule or process of interest, linking experimental data with available literature, and various other tasks which require aggregation of knowledge from a large number of scientific articles.

Among the main directions currently pursued within the BioNLP community is *event extraction*. This task involves the identification of biologically relevant events in scientific literature, covering both physical events involving genes and proteins as well as recursively defined regulatory events. Event arguments can have various semantic roles such as *cause* (effector) and *theme* (effectee). Event extraction was popularized through the BioNLP'09 and '11 Shared Tasks on Event Extraction (Kim et al., 2009; Kim et al., 2011), which allowed for a community-wide evaluation of numerous approaches to event extraction in a tightly controlled setting. The main advantage of the event representation is the relatively general and easily extensible definition of the task, as well as the level of detail provided for subsequent applications. An example of an event as defined in the Shared Tasks, is shown in Figure 1, illustrating *event nesting*, a crucial property of the event representation as well as the ability of events to abstract from the variation in natural language whereby a single event may represent a number of textually diverse statements. Additional details on event representation are given in the review of Ananiadou et al. (2010).

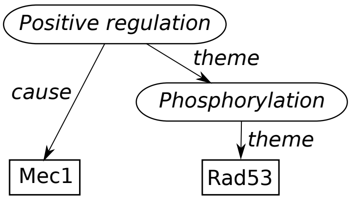


Figure 1: The event representation of the statements "*phosphorylation of Rad53 is controlled by Mec1*" (PMID:9315648), "*Mec1-dependent Rad53 phosphorylation*" (PMID:10449414), and "*... adaptor that enables Rad53 phosphorylation by Mec1*" (PMID:16085488).

As a follow-up to the BioNLP'09 Shared Task, the winning Turku Event Extraction System (TEES) (Björne et al., 2009) was applied to all PubMed abstracts, and the resulting set of 19 million extracted events was made publicly available for further research (Björne et al., 2010). This dataset was subsequently extended with event generalizations based on gene families and released as a relational database and web application[1] under the name *EVEX* by Van Landeghem et al. (2011; 2012).

The EVEX dataset addresses a fundamental shortcoming of the original event set: event extraction as defined in the Shared Tasks is purely text-based, i.e., the event extraction systems are not required to assign any biologically relevant identity (such as an Entrez Gene identifier) to the genes and proteins participating in the events. It is thus not possible to directly correlate the extracted events with other biological data, due to the numerous well-known issues caused by gene/protein name ambiguity (Chen et al., 2005). The EVEX dataset resolves this issue by assigning gene/protein mentions to their respective gene families; groups of homologous genes sharing sequence similarity. Gene families are retrieved from the publicly available resource Ensembl Genomes (Kersey et al., 2010) and every gene/protein mention in EVEX is assigned to at most one family. EVEX can thus define events with gene families as their arguments, rather than individual gene/protein mentions identified merely as character strings. Such events defined on top of entire families are referred to as *generalized*. For instance, the resulting family generalization of the event depicted in Figure 1 would have as its arguments the families *ATR* and *Rad53* with homologs in ca. 20 vertebrates, including human, and mouse.

The main advantage of the family generalizations is the fact that they rather straightforwardly support homology-based predictions, as sequence similarity often implies functional similarity. For example, if EVEX contains several regulatory events between pairs of genes that belong to families F_1 and F_2, this may be taken as supporting the hypothesis that other gene pairs belonging to F_1 and F_2 may exhibit a similar regulation pattern. Or, taken from a different perspective, given a pair of genes/proteins that are, based on experimental data, hypothesized to be involved in a regulation, the EVEX event generalizations can be used to straightforwardly access events among not only the given pair of genes/proteins, but also among their homologs.

In this study, we apply the EVEX dataset, as a literature-scale text mining resource, to the concrete BioNLP use case of identifying candidate regulators of NADP(H) metabolism in *Escherichia coli*. The purpose of this study is two-fold: First, we demonstrate the application of literature-scale event extraction to hypothesis generation in a real-life setting, driven by ongoing biological research on a specific molecule. Second, we aim at evaluating EVEX, and to some extent event extraction in general, to gain insight into its suitability for such hypothesis generation and to identify problem areas warranting future research.

2. Biological Motivation and Problem Setting

NADP(H) is a ubiquitous molecule that has a global role and the regulation of its metabolism is regarded as an ideal case-study in the well-studied model organism *E. coli*. NADP(H) is oxidized in more than 100 reactions while only three reactions contribute to the reduction of NADP$^+$, catalyzed by Zwf (Gdh), PntAB and Icd. The intracellular ratio of NADP(H) (reduced) to NADP$^+$ (oxidized) is tightly regulated under “normal” conditions (metabolic homeostasis) but is able to respond rapidly to changes in the intracellular environment, e.g. in the presence of reactive oxygen species (Ralser et al., 2007). NADP(H)-homeostasis is otherwise maintained at the border of thermodynamic limitations for whole cell-metabolism (Henry et al., 2007), with consequences for biotechnological applications (Walton and Stewart, 2004).

Whilst the regulation of the dynamic response is relatively well-established (*soxRS* regulon), there is currently no understanding of how NADP(H)-homeostasis is regulated in the absence of oxidative stress (Krapp et al., 2011). This study aims at identifying candidate regulators and other genes directly relevant to NADP(H)-homeostasis.

In a first step, genes that are known to influence NADP(H)-metabolism (typically enzymes or global regulators) were used to construct an initial list of *key genes* (KGs). This list was extended with *soxS/soxR* and *rob/marA*, well-studied genes that play a major role in the regulation of superoxide defense systems, as they are also known to influence the dynamic NADP(H)-response that is mediated by Zwf (Blanchard et al., 2007). Additional key genes were collected from EcoCYC (Keseler et al., 2011) and STRING databases (Jensen et al., 2009), leading to a final list of 14 key genes relevant to NADP(H)-metabolism that constitute the starting point for the text mining part of this study.

3. Related Work

The challenge of retrieving upstream regulators for any of the 14 key genes can be tackled by either querying *E. coli* specific knowledge bases, or by analyzing available literature.

3.1. *E. coli* Resources

PortEco (formerly EcoliHub) is a resource for laboratory strains of *E. coli*, providing a comprehensive summary on a queried gene by integrating data from EcoCYC (Keseler et al., 2011), EcoGene (Rudd, 2000), STRING (Jensen et al., 2009) and EcoliWiki (McIntosh et al., 2011). EcID (Andres Leon et al., 2009) further contains interactions extracted from KEGG (Kanehisa and Goto, 2000), MINT (Zanzoni et al., 2002) and IntAct (Hermjakob et al., 2004).

While these data sources provide valuable information on specific genes, the retrieved summaries sometimes lack pointers to experimental evidence, or merely link to full-text articles, preventing a quick manual validation of the results. Furthermore, the exponential growth of available experimental data in the life sciences prevents these resources from being fully up-to-date. Finally, organism-specific resources often exclude the retrieval of homology-based predictions. For these reasons, our aim was to track down candidate KG-regulators specifically from literature statements.

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

Search	**PubMed**		**Textpresso**	
	E. coli	**Any org.**	**EcoliWiki**	**EcoCyc**
NADPH	3,796	57,357	1,275	2,176
arcA	279	933	444	1,054
fnr	626	1,342	757	1,722
fruR	55	77	126	232
icd	50	16,005	132	388
marA	181	1,072	281	1,251
marR	139	1,905	310	1,213
pgi	103	2,069	165	414
pntA	8	13	25	75
pntB	8	11	25	37
rob	93	1,967	238	428
soxR	168	207	256	822
soxS	240	279	298	623
sthA	4	28	5	17
zwf	71	144	354	
Any KG	1,545	25,498	-	-
All articles	289,684	21,000,000	24,000	30,000

Table 1: Number of hits when searching for NADP(H) or the key genes in PubMed (with or without restricting the search to *E. coli*) or Textpresso (as implemented by EcoliWiki and EcoCyc).

3.2. Literature Search

Table 1 enumerates the number of articles retrieved from PubMed (Wheeler et al., 2007) when searching for one of the key genes in either *E. coli* or any organism. Further, it presents the results of querying the indexing framework Textpresso (Müller et al., 2004), as implemented by EcoliWiki or EcoCyc.

The large number of citations relevant to the key genes illustrates the necessity of fully automated text mining algorithms to manage the data abundance in the life sciences. For this purpose, many resources have previously been developed. For instance, iHOP allows fast retrieval of various relevant sentences for a certain gene, highlighting gene symbols, organism mentions and MeSH terms found within the same sentence (Hoffmann and Valencia, 2004). EBIMed covers Gene Ontology terms such as biological processes, as well as drugs and species names (Rebholz-Schuhmann et al., 2007).

The STRING database is a widely used resource containing protein-protein interactions predicted from text, amongst other resources (Jensen et al., 2009). The textual evidence is based on co-occurrence methods. PIE *the search* also searches PubMed for protein interaction data, using a classifier relying on word and syntactic features of whole articles (Kim et al., 2012).

For the case-study described in Section 2, we aim at retrieving more complex event structures, including various physical event types and regulatory events (cf. Figure 1). While the Medie search engine (Ohta et al., 2006) supports similar advanced queries, we focus specifically on the recently released EVEX resource (Van Landeghem et al., 2011), because its unique event family generalizations allow cross-species hypothesis generation, expanding the search domain also to homologs of the 14 key genes.

4. Methods and Resources

4.1. The EVEX Dataset

In this work, we use an extended version of the EVEX dataset, containing gene normalizations provided by the GenNorm system of Wei et al. (2011). The task of gene normalization is to disambiguate the gene and protein mentions in text to the biological object they represent, in our case by assigning them with a unique Entrez Gene identifier. The GenNorm system represents the state-of-the-art in gene normalization, having achieved first rank by several evaluation criteria in the BioCreative III Challenge (Lu et al., 2011). We used the Entrez Gene identifiers given by GenNorm to directly assign the gene and protein mentions to their corresponding Ensembl Genomes families. Where GenNorm does not assign an Entrez Gene identifier, the original algorithm of Van Landeghem et al. is used as a fallback.

Further, the dataset used in this study was also extended with events extracted from all full-text articles available in the Open Access subset of PubMed Central, substantially increasing the amount of literature available for text mining. The impact of this extension is separately evaluated in Section 5.1.

4.2. EVEX Event Preprocessing

As illustrated in Figure 1, events may constitute complex structures where an event may have as its argument another, recursively nested event. While these structures properly account for the semantics of the underlying natural language statements, they cannot be directly correlated with the vast majority of existing biological resources which generally take the form of networks of pairwise interactions between genes and proteins.

To this end, we have defined a rule-based procedure to decompose complex events into pairwise directed interactions. (Van Landeghem et al., 2012) This procedure assigns three interaction types: *regulation* (directed), *indirect regulation* (directed), and *binding* (undirected), stemming from the fact that only regulation and binding events may have more than one argument in the event scheme defined by the Shared Tasks and therefore can generate pairwise interactions. In this work, we further merge *regulation* and *indirect regulation* into a single *regulation* type. The result of applying this procedure to the common event structure of one gene regulating the interaction of two other genes, is shown in Figure 2. Since much of this study deals with generalized events, i.e., events defined on top of gene families, the result of the decomposition procedure will correspondingly be pairs of gene families.

The procedure may at first seem to be defeating the purpose of defining and extracting detailed event structures, since, as illustrated in Figure 2, the event representation captures the semantics of the underlying statement more accurately than the extracted pairs. However, it must be noted that in our current application setting, the underlying events are preserved: the pairwise interactions are used to identify events of interest, which are subsequently presented in full detail to the end-user. Therefore, rather than redefining events per se, we are merely defining a layer of simplified, pairwise interactions on top of the events. This layer serves as an interface between the events in EVEX and pairwise biological data, such as microarray co-expression studies as well as other existing and widely used resources.

4.3. EVEX Candidates

To search for novel regulators of the 14 given *key genes* from *Escherichia coli* strain *K-12* substrain *MG1655*, we first determine their families in Ensembl Genomes, resulting in 14 *key families* (KFs). Next, we extract all events from EVEX which involve at least one key family, regardless of its role in the event (cause or theme). The search is performed on the level of family pairs, as described in Section 4.2. We therefore obtain pairs of the following three types: *Binding(X,KF)*, *Regulation(X,KF)* and *Regulation(KF,X)*, where *X* is a *candidate family* of interest. Since the problem setting is specific to the aforementioned substrain of *E. coli*, we discard all events where the candidate family *X* does not contain a gene from our target organism. Subsequently, we manually evaluate all these extracted events, only preserving correctly extracted or otherwise biologically relevant events where both arguments are assigned to their correct family. The results of this manual evaluation are presented in detail in Section 5.1.

The final set of events that were evaluated as fully correct comprised 132 event occurrences of 81 unique Ensembl Genomes generalized events. These events linked 41 unique candidate gene families to 12 of the initial key gene families. For two key gene families, no EVEX events were found. From this final set of events, we constructed an *E. coli*-specific gene network (referred in further text as the EVEX network) by selecting the *E. coli* gene member in each family. The network is shown in Figure 3.

4.4. Microarray Data

Microarray data was collected from the Affymetrix chip [Ecoli_Asv2] (Affymetrix *E. coli* Antisense Genome Array). Specific microarray data were selected based on their expected relevance for NADP(H)-metabolism in *E. coli*. In order to ensure consistency across all treatments, all data was extracted from only two extensive series of microarray analyzes carried out by Covert et al. (2004), focusing on oxidative stress, and Dong et al. (2008), focusing on the global regulator RpoS. The transcriptome data were extracted from Gene Expression Omnibus (GEO) with accession number GPL199 (Barrett et al., 2011). The microarray analysis platform Chipster was used to generate normalized expression values and p-values. Networks were constructed, analyzed and visualized with the freely available software Cytoscape (Shannon et al., 2003). Plug-in ExpressionCorrelation[2] was employed to construct networks using expression and significance values. A similarity network strength threshold of 0.65 was selected to calculate a similarity matrix using the Pearson correlation coefficient.

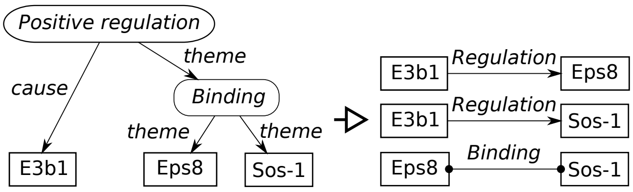


Figure 2: Pairwise decomposition of an event with recursive nesting from the statement "*E3b1 (. . .) plays a critical role (. . .) by facilitating the interaction of Eps8 with Sos-1*" (PMID:15178460).

The co-expression based gene network (referred to as CoEx) thus obtained was then overlaid with the EVEX network, as shown in Figure 3.

5. Results and Discussion

In the following section, we analyze and discuss the results from two perspectives: evaluation of the text mining methods and resources used, and the biological relevance of the findings.

5.1. Text Mining Findings

The set of initial candidate events extracted from EVEX comprises of 348 unique generalized events aggregated from 461 individual event occurrences in text. Each of these events, by definition, has at least one key family as an argument. In total, these events involve 152 unique families. In the following, we manually evaluate the 461 candidate event occurrences based on two criteria: correctness of the extracted event, i.e., whether the event reflects the statement from which it was extracted and, as a second criterion, the correctness of the assignment of the gene and protein mentions to their respective families. This second criterion is particularly crucial when using the events for family-based hypothesis generation.

There were 243 (53%) correctly extracted event occurrences comprising 169 unique generalized events, well in line with the precision figures reported for the Turku Event Extraction System in the official BioNLP'09 Shared Task evaluation for multiple-argument events (50% for bindings and 46% for regulations) (Kim et al., 2009). In addition to genuine false positives, we found among the remaining 218 events two classes of events which, although considered false positives from the strict event definition point of view, were judged biologically relevant and were thus considered for further evaluation. First, these include 36 relevant events extracted with incorrect type, either through label substitution of regulation vs. binding, or constituting a relationship which does not have an appropriate type defined in the event representation. Secondly, three events were found encoding regulation in the opposite direction (i.e. rather than the correct *Regulation(X,KF)*, the false positive *Regulation(KF,X)* is extracted).

Since we are interested in events which recover upstream regulators and binding partners of the key families, we disregard from further evaluation all events of the type *Regulation(KF,X)* (naturally, still preserving *Regulation(KF,KF)*). The remaining 183 events (representing 118 unique generalized events and 76 unique families), comprising both true positives and corrected relevant false positives, were evaluated for the correctness of gene family assignment. Of these, 132 (72%) events were such that both arguments were resolved to their correct Ensembl Genomes family in EVEX. These fully-correct occurrences constitute 81

[2] http://chianti.ucsd.edu

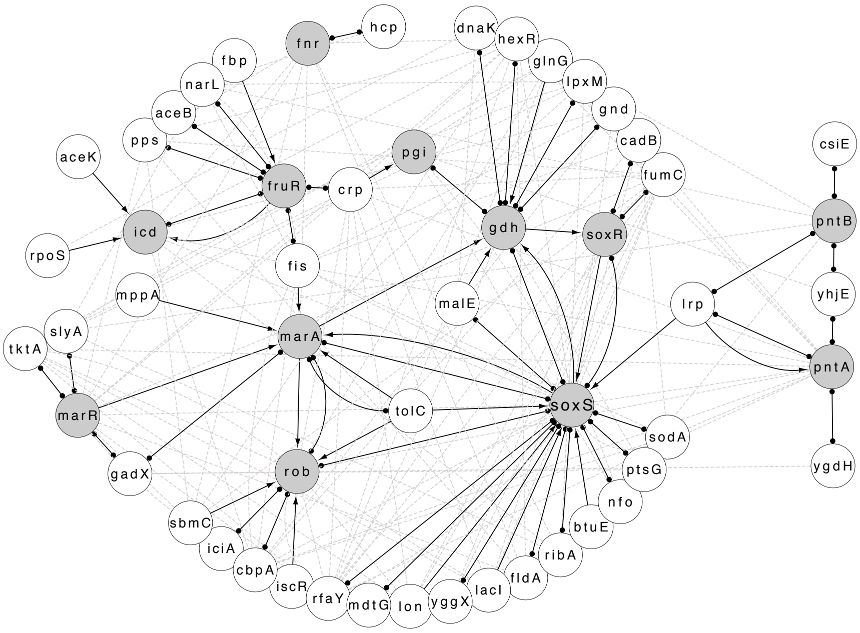


Figure 3: The complete network obtained from EVEX (solid lines) and microarray-based co-expression analysis (dashed lines). In the EVEX network, circle-terminated connections indicate binding and arrows indicate regulation. The key genes are highlighted in gray; 2 key genes are not present since no EVEX events were extracted for them. Note that only events involving at least one key gene are extracted, therefore no events between candidate genes are present.

unique generalized events and 53 unique families (12 key families and 41 candidate families).

To summarize, the precision of the two key components is 53% for event extraction and 72% for gene family assignment of both arguments. However, since the errors are cumulative, the overall precision of even state-of-the-art systems leaves room for improvement. In addition, it is also important to note that a number of false positives among the events do bear biological significance and were deemed relevant for the current study. Naturally, this is highly use case specific and should not be interpreted as an attempt to artificially boost the precision figures.

Manually evaluating the initial set of events to construct the EVEX network amounted to a little less than three days of work of one person. Of the two validation steps (event correctness and family assignment), evaluating the correctness of the family assignments was clearly the more labor-intensive one, as it often required careful identification of the species, strain, and sub-strain involved — information rarely present in the abstract. However, in order to be able to rely on the integration of the EVEX and CoEx networks, we consider the manual evaluation step of great importance and not excessively labor-intensive, particularly compared to the effort that would be necessary to build such a network without any text mining support.

Finally, we discuss the issue of event extraction from full-text articles versus abstracts. The need for text mining in full text articles, in addition to PubMed abstracts, is becoming broadly recognized in the BioNLP community. The EVEX dataset, as used in this study, was extended with full-text articles from the PubMed Central Open-Access (PMC-OA) section and we can thus evaluate the impact of full-text mining on this real-world use case. We find that 18 out of the 41 candidate families, i.e. nearly half, were identified only from a body of a full-text article. This figure clearly demonstrates the added value of full-text articles for text mining and, consequently, the importance of opening full-text articles for automated access.

5.2. Biological Findings

In order to analyze the EVEX network relative to the CoEx network, we initially focus on three important patterns in which these two networks can support each other, illustrated in Figure 4.

Direct support for EVEX-identified relationships by the co-expression network (pattern A) was found in two cases: *sodA-soxS* and *soxS-rfaY*. Both were determined to be genuine positives based on detailed experimental evidence including chromatin immunoprecipitation, DNA-binding and co-expression. Further, 49 triangular clusters (24 of type B, 23 of type C, and 2 which can be classified as either B or C) were identified.

On one hand, such triangular relationships may indicate

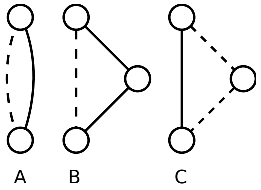


Figure 4: Three patterns of particular interest when referencing the EVEX (solid lines) and CoEx (dashed lines) networks. The two networks may fully support each other (A), the CoEx network may provide further support for an indirect relation from the EVEX network (B), or, finally, the EVEX network may provide further explanation for an indirect relation in the CoEx network (C).

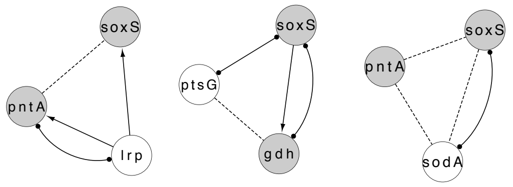


Figure 5: Two type B patterns and one type A/C pattern with varying biological explanations.

that apparent indirect interactions may in fact be direct. One such example is the type B pattern between *gdh*, *soxS*, and the gene encoding for a component of the central glucose-uptake system *ptsG*, shown in Figure 5 (middle). The pattern suggests that a link between *NADP(H)-metabolism* and glucose uptake may exist which warrants further experimental investigation.

On the other hand, a type B pattern may also indicate the very opposite: relationships which, even though appearing direct in one network, are in fact shown to be indirect, in light of the relationships present in the other network. Consider, for instance, the type B pattern observed between *pntA*, *soxS*, and *lrp* in Figure 5 (left). The *pntA–soxS* co-expression (a direct relationship in the CoEx network) is most likely an indirect relationship caused by co-regulation, since both *pntA* and *soxS* are members of the *lrp* regulon. This becomes apparent from the EVEX network, and the actual statements underlying the EVEX events.

An example of a type C pattern (which also contains a type A sub-pattern) is illustrated in Figure 5 (right). One EVEX binding event was identified between the transcriptional regulator *soxS* and *sodA* encoding superoxide dismutase. Both of these two genes are known to respond and contribute to alleviate oxidative stress, whilst *pntA* until now is not known to be involved in such a metabolic manner. A connection between all three genes was identified in CoEx, supporting the EVEX event and also interestingly linking PntAB to dynamic stress conditions which until now it has not been described to be involved in.

These three examples serve to illustrate the diversity of hypotheses obtained from an initial analysis of simple triangular patterns in the combined EVEX/CoEx networks.

The ability to support homology-based function prediction has been presented as one of the primary motivations for the family-based generalization in EVEX. Therefore, candidate genes identified from organisms other than *E. coli* warrant a closer inspection. Of the 41 candidate genes, only five originated entirely from non-*E. coli* studies and further three originated both from *E. coli* and non-*E. coli* literature. In total, about 20% of candidate genes were thus identified through the generalization. This relatively low number is likely due to *E. coli* serving as a model organism in studies of prokaryote central carbon metabolism — the effect of the generalization would likely be more pronounced if the target organism was less studied, or the gene families more coarsely defined, in which case candidates from a related model organism would be identified through the generalization. The five candidates extracted uniquely from non-*E. coli* literature are summarized in Table 2, together with the number of co-expression associations with the key genes. *hexR* is clearly the most “interconnected” candidate and we thus select it for further discussion. *hexR* has not previously been studied in *E. coli* but has been shown to act as a transcriptional regulator of several genes encoding enzymes in central carbon metabolism of *Pseudomonas putida*, including the NADP$^+$-reducing glucose-6-phosphate dehydrogenase (gdh) (Daddaoua et al., 2009). Interestingly, *hexR* is located adjacent to *gdh* in the genome of *E. coli* and they both appear to share the promoter-region with binding-motifs for SoxS, MarA and Rob. There is no direct co-expression between *gdh* and *hexR*, however, triangular type B patterns are observed with both *gnd* and *soxS* (Figure 6). The microarray-analysis also shows co-expression with both *icd* and *pntA*, closely linking *hexR* with all three NADP$^+$-reducing enzymes. *hexR* therefore represents a highly interesting candidate to study further.

EVEX event	# of co-expressed KGs
hexR - gdh	4
glnG - gdh	2
cadB - gdh	2
lpxM - gdh	2
slyA - marR	1

Table 2: The number of key genes co-expressed with the candidate genes identified by EVEX in organisms other than *E. coli*.

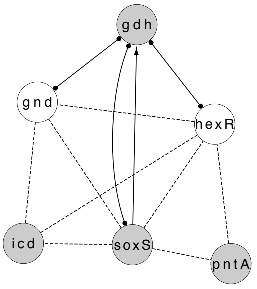


Figure 6: *HexR*-related sub-network in EVEX and in CoEx.

In summary, several new relationships were uncovered by the combined analysis generating several testable and potentially interesting hypotheses, in particular the notion that NADP$^+$-reduction is subject to coordinated regulation by the transcriptional regulators SoxS and HexR (Figure 6). Importantly, even though not previously studied in the target host organism, *hexR* could be identified by EVEX-analysis alone and further supported by triangular relationships involving also CoEx. The relative lack of EVEX-events for the critical transhydrogenases, despite a wealth of edges in CoEx, supports our conclusion that regulatory interactions influencing non-dynamic NADP(H)-homeostasis still remain to be explored in prokaryotes.

6. Conclusions and Future Work

We have demonstrated the application of EVEX, a literature-scale event extraction resource to a real-world biological use case, with an encouraging result. With a reasonable manual effort, we were able to extract a network of candidate genes related to the metabolism of NADP(H) in *E. coli*, starting with 14 key genes, and to integrate the network with microarray-based co-expression data. Integrating the two networks and using them as mutually supporting resources was a crucial step, and we were able to identify several candidate genes of particular interest, warranting further experimental evaluation. This study was only possible because the predictions of the event extraction system could be, via the gene family assignment procedure implemented in EVEX, directly related to available experimental data, focusing specifically on genes from the target organism, or their homologs.

Our evaluation has shown that, even when state-of-the-art event extraction and gene normalization systems are employed, automatically extracted text mining results need further manual validation to enable meaningful integration with experimental data in similar focused use cases. However, we expect that after this initial case study the manual effort involved in the process can be further decreased by developing tools specifically supporting such applications, for instance focusing on the labor-intensive task of gene family assignment evaluation.

Since NADP(H) is a metabolite, and not a gene/protein, it falls out of scope in the majority of BioNLP studies. Metabolites are of great relevance and it is important to focus on incorporating events pertaining to metabolites into the EVEX dataset. This can be supported by the methods developed for the BioNLP'11 Shared Task (Kim et al., 2011) that involved metabolites in the ID sub-task. A further challenge is presented by the fact that metabolites cannot be assigned into a gene family, which is a strict requirement of the current event generalization procedure in EVEX. A more relaxed criterion will therefore need to be implemented so as to account for events among different classes of bio-entities, most importantly between proteins and metabolites, without losing the benefit obtained from the family-based generalization.

Further future work can be charted in several directions. First, the current network can be expanded by extracting and verifying events among the currently identified candidate families, as well as including events directly involving NADP(H). Then, the network can be expanded by binding partners and regulators of the current candidate families, essentially adding a layer of 2^{nd} degree regulators. Since the network is expected to grow substantially and manual evaluation of all 2^{nd} degree regulators may not be feasible, it will be important to investigate external resources as well as internal statistics which can be used to rank the new candidates and focus the exploration of the network to the most promising areas. Finally, since the use case presented in this study is an example of what we expect to be a commonly faced problem, we will consider developing novel tools to support and automatize building the network without requiring extensive data-processing skills.

7. Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement n. [260661] (P.R.J.) and Academy of Finland projects no. 253269 (S.Kr.) and 138796 (S.Ka.). S.VL. and Y.vdP. further acknowledge the support of Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks"). S.VL. also thanks the FWO (research foundation Flanders) for funding her research.

8. References

S. Ananiadou, S. Pyysalo, J. Tsujii, and D.B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*, 28(7):381–390.

E. Andres Leon, I. Ezkurdia, B. García, A. Valencia, and D. Juan. 2009. EcID. a database for the inference of functional interactions in E. coli. *Nucleic Acids Research*, 37(suppl 1):D629–D635.

T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. 2011. NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic acids research*, 39(suppl 1):D1005.

J. Björne, J. Heimonen, F. Ginter, A. Airola, T. Pahikkala, and T. Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 10–18, Morristown, NJ, USA. Association for Computational Linguistics.

J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, and T. Salakoski. 2010. Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the BioNLP 2010 Workshop*, pages 28–36. Association for Computational Linguistics.

J.L. Blanchard, W.Y. Wholey, E.M. Conlon, and P.J. Pomposiello. 2007. Rapid changes in gene expression dynamics in response to superoxide reveal SoxRS-dependent and independent transcriptional networks. *PLoS ONE*, 2(11):e1186, 11.

L. Chen, H. Liu, and C. Friedman. 2005. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21:248–256.

M.W. Covert, E.M. Knight, J.L. Reed, M.J. Herrgard, and B.O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96.

A. Daddaoua, T. Krell, and J.L. Ramos. 2009. Regulation of glucose metabolism in Pseudomonas: The phosphorylative branch and Entner-Doudoroff enzymes are regulated by a repressor containing a sugar isomerase domain. *Journal of Biological Chemistry*, 284(32):21360.

T. Dong, M.G. Kirchhof, and H.E. Schellhorn. 2008. Rpos regulation of gene expression during exponential growth of Escherichia coli K12. *Molecular Genetics and Genomics*, 279:267–277. 10.1007/s00438-007-0311-4.

C.S. Henry, L.J. Broadbelt, and V. Hatzimanikatis. 2007. Thermodynamics-based metabolic flux analysis. *Biophysical journal*, 92(5):1792–1805.

H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler. 2004. IntAct: an open source molecular interaction database. *Nucleic Acids Research*, 32(suppl 1):D452–D455.

R. Hoffmann and A. Valencia. 2004. A gene network for navigating the literature. *Nat Genet*, 36(7):664, Jul.

L.J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering. 2009. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research*, 37(suppl 1):D412–D416.

M. Kanehisa and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28:27–30.

P.J. Kersey, D. Lawson, E. Birney, P.S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kähäri, R. J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A.J. Vilella, and A. Yates. 2010. Ensembl Genomes: Extending Ensembl across the taxonomic space. *Nucleic Acids Research*, 38(suppl 1):D563–D569.

I.M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A.G. Shearer, A. Mackie, I. Paulsen, R.P. Gunsalus, and P.D. Karp. 2011. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, 39(suppl 1):D583.

J.D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 1–9, Morristown, NJ, USA. Association for Computational Linguistics.

J.D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics.

S. Kim, D. Kwon, S.Y. Shin, and W.J. Wilbur. 2012. PIE the search: searching PubMed literature for protein interaction information. *Bioinformatics*, 28(4):597–598.

A.R. Krapp, M.V. Humbert, and N. Carrillo. 2011. The soxRS response of Escherichia coli can be induced in the absence of oxidative stress and oxygen by modulation of NADPH content. *Microbiology*, 157(4):957.

Z. Lu, H.Y. Kao, C.H. Wei, M. Huang, J. Liu, C.J. Kuo, C.N. Hsu, R.T. Tsai, H.J. Dai, N. Okazaki, H.C. Cho, M. Gerner, I. Solt, S. Agarwal, F. Liu, D. Vishnyakova, P. Ruch, M. Romacker, F. Rinaldi, S. Bhattacharya, P. Srinivasan, H. Liu, M. Torii, S. Matos, D. Campos, K. Verspoor, K.M. Livingston, and W.J. Wilbur. 2011. The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12(Suppl 8):S2+.

B.K. McIntosh, D.P. Renfro, G.S. Knapp, C.R. Lairikyengbam, N.M. Liles, L. Niu, A.M. Supak, A. Venkatraman, A.E. Zweifel, D.A. Siegele, and J.C. Hu. 2011. EcoliWiki: a wiki-based community resource for Escherichia coli. *Nucleic Acids Research*.

H.M. Müller, E.E. Kenny, and P.W. Sternberg. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, 09.

T. Ohta, Y. Miyao, T. Ninomiya, Y. Tsuruoka, A. Yakushiji, K. Masuda, J. Takeuchi, K. Yoshida, T. Hara, J.D. Kim, Y. Tateisi, and J. Tsujii. 2006. An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 17–20. Association for Computational Linguistics.

M. Ralser, M.M. Wamelink, A. Kowald, B. Gerisch, G. Heeren, E.A. Struys, E. Klipp, C. Jakobs, M. Breitenbach, H. Lehrach, and S. Krobitsch. 2007. Dynamic rerouting of the carbohydrate flux is key to counteracting oxidative stress. *Journal of biology*, 6(4):10.

D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. 2007. EBIMed–text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):e237–e244.

K.E. Rudd. 2000. EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Research*, 28(1):60–64.

P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.

S. Van Landeghem, F. Ginter, Y. Van de Peer, and T. Salakoski. 2011. EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of the BioNLP 2011 Workshop*, pages 28–37. Association for Computational Linguistics.

S. Van Landeghem, K. Hakala, S. Rönnqvist, T. Salakoski, Y. Van de Peer, and F. Ginter. 2012. Exploring biomolecular literature with EVEX: Connecting genes through events, homology and indirect associations. *Advances in Bioinformatics*. To appear.

A.Z. Walton and J.D. Stewart. 2004. Understanding and improving NADPH-dependent reactions by non-growing Escherichia coli cells. *Biotechnology Progress*, 20(2):403–411.

C.H. Wei and H.Y. Kao. 2011. Cross-species gene normalization by species inference. *BMC bioinformatics*, 12(Suppl 8):S5.

D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko. 2007. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35(suppl 1):D5–D12.

A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. 2002. MINT: a Molecular INTeraction database. *FEBS Letters*, 513(1):135 – 140.

Annotating and Evaluating Text for Stem Cell Research

Mariana Neves[1], Alexander Damaschun[2], Andreas Kurtz[2], Ulf Leser[1]

[1]Humboldt-Universität zu Berlin, Knowledge Management in Bioinformatics, Berlin, Germany,
[2]Berlin Brandenburg Center for Regenerative Therapies, Charité, Berlin, Germany
neves@informatik.hu-berlin.de, alexander.damaschun@charite.de,
andreas.kurtz@charite.de, leser@informatik.hu-berlin.de

Abstract

The regeneration of vital organs and tissues remains one of the biggest medical challenges. However, the use of embryonic stem cells and induced pluripotent stem cells allows novel replacement strategies. The CellFinder project aims to create a stem cell data repository by linking information from existing public databases and by performing text mining on the research literature. We present the first version of our corpus which is composed of 10 full text documents containing more than 2,100 sentences, 65,000 tokens and 5,200 annotations for entities. The corpus has been annotated with six types of entities (anatomical parts, cell components, cell lines, cell types, genes/protein and species) with an overall inter-annotator agreement around 80%. Preliminary results using baseline methods based on freely available terminologies and systems have returned a recall which ranges from 48% to 90% for the extraction of the named entities. The high distribution of entities which are representative of the stem cell research, specially cell types, makes our corpus a valuable resource for the stem cell domain.

Keywords: biomedical corpus, names-entity recognition, stem cell research.

1. Introduction

The regeneration of lost vital organ and tissue function after severe injury or end-stage progression of diseases remains one of the biggest unmet medical challenges (Viswanathan and Keating, 2011). Despite pharmacological advances in alleviating the symptoms of compromised vital functions or in slowing disease progression, the only available therapy for permanent impairment or organ loss is organ replacement. However, since there are few indications for which sufficient numbers of donors exist in order to meet the demand for transplant organs (Watson and Dark, 2012), alternative strategies are needed to restore organ and tissue function.

The advent of human embryonic stems cell (hESCs) (Thomson et al., 1998) and human induced pluripotent stem cells (hiPSCs) (Yu et al., 2007) together with the identification of many types of multipotent precursor and adult stem cell (Barile et al., 2011) have opened promising new routes for novel replacement strategies (Atala, 2012). Some of these approaches aim to activate the body's endogenous regenerative capacities, others look at the stem cells' capacity to differentiate into specific cell types for direct application in cell therapy or use them as building blocks in tissue engineering.

All regenerative approaches involving stem cells or their differentiated progeny have one fundamental requirement in common: The cells to be used have to be both effective and safe. Therefore, therapeutic cell populations to be applied in the patient or in engineered tissue have to be well characterized based on reliable measurement and analysis techniques as well as validated by knowledge bases of stem cells and their progeny (Wohlers et al., 2009; Kerrigan and Nims, 2011).

Results of such studies create an ever-rising flood of scientific information and experimental data that is virtually impossible to be registered, analyzed or exploited without the aid of sophisticated bioinformatics applications running on powerful computer infrastructures. This is particularly evident in the rising field of regenerative medicine, in which several specialized scientific disciplines are combined (Viswanathan and Keating, 2011). Alongside clinical cell-directed pathological and cytological data, additional information such as cell-anatomical, cell-biological, genetic and biochemical data as well as potencies and functional interactions are required for the modeling, prediction and analysis of cell-based therapies, as well as for basic cell research. Consequently regenerative medicine will not progress without an integrating, systematic and analytical approach that utilizes adequate shared information and data resources (Hatano et al., 2011; Jung et al., 2010).

CellFinder[1] is based on the idea of establishing a central stem cell data repository, by utilizing and interlinking existing public databases regarding defined areas of human pluripotent stem cell research. Provision of standardized description, registration and interlinking of stem cell data on the above mentioned levels is a prerequisite for the effective exchange of data. One specific aim of CellFinder is to identify processes by which various kinds of stem and precursor cells may differentiate, function and react and subsequently be applied. An important source of knowledge are published research results. In CellFinder, text mining methods are employed to extract knowledge from this scientific literature, which will be further made available in our on-line repository.

In the last years, we have observed an increase in the availability of corpora for the biomedical domain (Kim et al., 2003; Pyysalo et al., 2008). In the last 10 years, the biomedical natural language community has migrated from sentences-based corpora annotated with one or a couple of

[1] http://www.cell-finder.org/

Sections	Sentences	Tokens	Annotations						
			Anatomy	Cell Comp.	Cell Line	Cell Type	Gene	Species	TOTAL
Abstract	79	2683	88	6	10	151	45	24	324
Introduction	225	6881	155	17	7	302	56	59	596
Methods	539	15540	130	64	101	228	356	109	988
Results	1052	31975	423	91	187	832	1036	191	2760
Discussion	256	7221	99	20	15	245	112	47	538
Conclusion	26	731	18	0	8	19	16	8	69
TOTAL	2177	65031	913	198	328	1777	1621	438	5275

Table 1: Number of sentences, tokens and annotations per entity and per section in the full text document. A total per type of entity and per sections is shown in the last line and last column, respectively.

named-entities (Rosario and Hearst, 2004; Tanabe et al., 2005) to the annotation of abstracts with more than one type of entity (Kim et al., 2003; Klinger et al., 2008; Furlong et al., 2008), relationships between entities (Pyysalo et al., 2008) and biological events (Kim et al., 2008). More recently, also full texts have become popular (Kim et al., 2011; Carreira et al., 2011). Finally, the community-based effort for the construction of the CALBC silver standard corpus composed by a variety of entities (Rebholz-Schuhmann et al., 2010) is certainly helpful for the biomedical natural language processing research.

Studies have shown that the structures of abstract and full text are different (Cohen et al., 2010) and that more valuable information is usually found only in full texts. We have indeed noticed that the data which is relevant for the CellFinder's database is usually present only in the results sections of the publications. Therefore, in order to support the development and evaluation of our text mining methods, some selected full text documents have been annotated with entities and biological processes relevant for the stem cell domain. The annotation schema includes a variety of entities, such as cell lines, anatomical parts and genes/proteins, as well as biological events, such as gene expression and differentiation. We present here the first version of our corpus which is composed of 10 full text documents comprising 2,177 sentences, 65,031 tokens and 5,275 annotations of entities.

2. Overview of the Corpus

We present ongoing work which aims at annotating a corpus on the stem cell domain with semantic entities, biological events as well as associated meta-knowledge (Thompson et al., 2011). Our annotation schema consists of six types of entities:

- anatomical parts (i.e., tissues, organs and body parts): "bone marrow", "adipose tissue";
- cell components: "membrane", "chromosome", "nuclei";
- cell lines: "hESMPC9.1", "H1";
- cell types: "mesenchymal precursors", "skeletal muscle cells";
- genes/proteins: "OCT4", "vimentin";
- species: "human", "mouse".

The importance of each of these entities in the stem cell research is evident. We now give a more detailed description of each of them.

Anatomical parts entities describe the spacio-temporal locations of cell types throughout their existence/development in tissues, organs (and part thereof), body parts and organisms. The annotation of species is necessary in order to map homologies between different organisms and to transfer insights from established animal models to the human organisms and vice versa. Anatomy has also been applied to in vitro anatomies formed by cells, e.g. embryoid bodies, monolayers or rosettes.

Cell component refers to sub-cellular structures or locations within a cell (sometimes specific to a certain cell type) where genetic functions are exerted, proteins are expressed or molecules are detected.

Cell lines describe instances of cells of a certain type that have been modified with biomolecular, genetic, chemical or physical techniques in order to preserve one or several properties of their specific type or to arrest the cells in a certain stage of their development. This enables the cells to be cultivated reproducibly over prolonged periods of time (compared with the naive status), or, in the case of immortalized cells, indefinitely. Designations for cell lines are commonly arbitrary and originate from their providers.

Cell type encompasses all instances of a biological cell (individual cells, colonies or agglomerations in biological tissue) with a distinct set of morphological, biomolecular and functional properties. With the exception of terminally differentiated adult/somatic cells, all cells of a certain type have at least one precursor cell type and at least one progeny cell type.

Genes or proteins refers to instances in the text that mention gene names and functions, RNA that has been transcribed from any particular gene or a protein that has been expressed as the result of gene (up)regulation. The same is true for any mention of the absence of a protein (or its expression) or the suppression of a gene.

Annotations have been performed by two experts from the stem cell domain. Annotator 1 is a biologist with extensive expertise in molecular and stem cell biology (mesenchymal stem cell, hESC, hiPSC), cell generation, characterization and GXP manufacture, systems biology, state-of-the-art analysis techniques, clinical studies and biomedical ethics. Annotator 2 is a biotechnologist with long-years expertise in stem cell characterization and registration, cell-based knowledge bases and dissemination and tissue engineering.

In this first round of annotations, 10 full text documents have been annotated. Papers have been selected based on the work of (Löser et al., 2010) in which publications on the field of human embryonic stem cells have been surveyed (up to November of 2009). A list of 990 publications have been derived from this work (available as supplementary material). From this list, 62 are included in the PubMed Central Open Access Subset, and thus, can be freely used for text mining purposes. Our annotators have selected 10 full papers for the annotation, namely PMIDs: 16316465, 17381551, 17389645, 18162134, 18286199, 15971941, 16623949, 16672070, 17288595 and 17967047. Full texts were obtained in XML format from the Pubmed Central Open Access Subset page[2]. For performing the annotations, we used Brat[3] (brat rapid annotation tool) (Stenetorp et al., 2012). The documents had to be split into sections due to the compromised performance of Brat when dealing with long documents. The number of annotations found for each entity type in the various sections of the full paper is shown in Table 1. An example of some of the annotations for the six entity types is shown in Figure 1.

The gold-standard corpus was created by merging the annotations from both annotators. An automatic consensus was carried out to remove overlapping annotations, such as singular and plural forms (e.g., "stem cell" and "stem cells") and mentions starting with hyphens (e.g., "-H1.3" and "H1.3"). Additionally, we automatically checked those mentions which started or ended with parenthesis, curly or squared brackets, which are certainly due to a mistake when selecting the text of the mention. Finally, the documents were manually checked and some few overlapping inconsistencies have been corrected, such as "mesenchym" and "mesenchymal", by keeping only the larger one.

Entities	Exact	Overlap+Type	Overlap
Anatomy	0.37	0.61	0.76
Cell Comp.	0.33	0.39	0.49
Cell Line	0.75	0.92	0.95
Cell Type	0.30	0.85	0.91
Gene/Protein	0.77	0.81	0.83
Species	0.78	0.81	0.83
TOTAL	0.51	0.80	0.85

Table 2: F-score of the inter-annotator agreement for each of the entities.

The inter-annotator agreement (IAA) was computed as F-score and is shown in Table 2. Annotations which matched exactly regarding the span and the type of entity are shown as "Exact". Overlapping annotations which belonged to the same type of entity were also included in the consensus corpus as alternative synonyms. The agreement when considering these cases is found in Table 2 under the column "Overlap+Type". For instance, one of the annotators identified "human embryonic stem cells" as a cell type, while the other annotated just "stem cells". Alternative synonyms are not unusual in the biomedical domain, such as the GENETAG corpus (Tanabe et al., 2005) which includes synonyms for gene and protein names. We have also decided to add overlapping annotations which belong to different types of entities. For instance, for the same example above, when one of the annotators identified "human embryonic stem cells" as a cell type, the other one annotated "human" as a species and "embryonic" as anatomical part. We show the increment in the IAA when allowing overlapping for annotations of different types in the "Overlap" column of Table 2. Finally, as an ongoing work, entities which have been annotated by only one of the annotators were integrated into the gold standard in this phase of the project.

The corpus was made available in our repository of corpora[4] in its full text version and also split by sections. For the visualization of the corpus, we recommend Chrome, Safari or Opera. The full text and the sections-split versions of the corpus are available for downloading from the corpus web page[5] in the standoff format used by Brat and in the XML format used in (Pyysalo et al., 2008).

3. Preliminary Evaluation

In this section, we present our preliminary results for predicting entities in the stem cell domain. Regarding the recognition of the entities annotated in this corpus, we are more concerned about the recall. If a certain entity cannot be found in the text during the named-entity recognition step, the events in which it participates will not be found either.

As baseline, we decided to use only dictionary-based methods derived from existing ontologies or terminologies and freely available systems. Thus, we did not use the annotated corpus to train a specific tagger for any of the entity types. Details for the methods used in the recognition of each entity type are presented below.

We used Metamap (Aronson and Lang, 2010) for extracting annotations for five entity types: anatomical parts, cell components, cell type, gene/protein and species. We restricted the annotations to certain semantic types using the "-J" parameter. The mapping of the semantic types to our entities is shown in Table 3. Additionally, we used the parameter which allows variants for acronyms and abbreviations ("-a"). As Metamap does not work properly with long texts, we split the full text documents into sentences using the sentence detector available in OpenNLP[6]. Additionally, also due to the inability of Metamap in processing long sentences, we only analyzed those under 1000 characters, resulting in five discarded sentences.

Anatomical parts and cell components were extracted only using Metamap. For cell types, besides Metamap, a dictionary of cell type synonyms was created using the OBO Cell Type ontology (Bard et al., 2005). We used Lingpipe

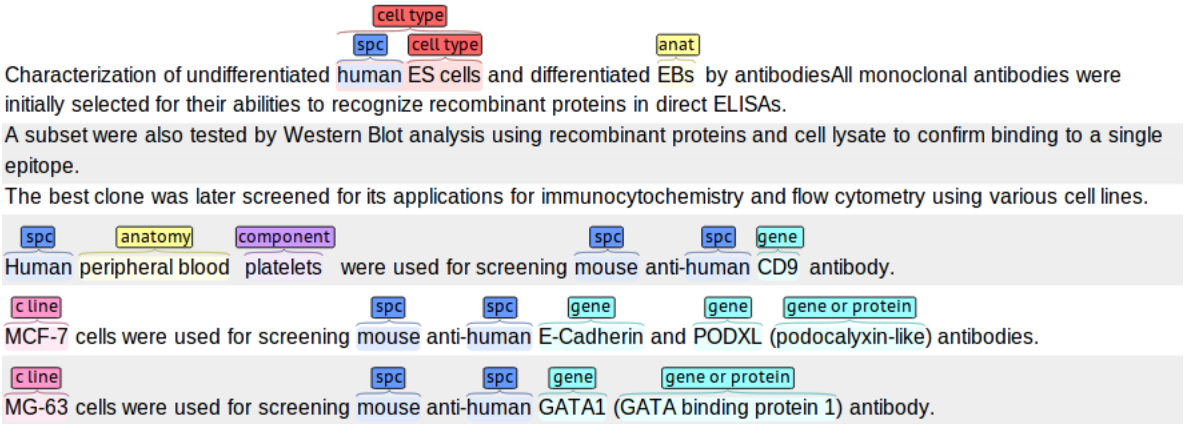
[2] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[3] http://brat.nlplab.org

[4] http://corpora.informatik.hu-berlin.de/index.xhtml

[5] http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/resources/cellfinder/

[6] http://incubator.apache.org/opennlp/



Characterization of undifferentiated human ES cells and differentiated EBs by antibodiesAll monoclonal antibodies were initially selected for their abilities to recognize recombinant proteins in direct ELISAs.

A subset were also tested by Western Blot analysis using recombinant proteins and cell lysate to confirm binding to a single epitope.

The best clone was later screened for its applications for immunocytochemistry and flow cytometry using various cell lines.

Human peripheral blood platelets were used for screening mouse anti-human CD9 antibody.

MCF-7 cells were used for screening mouse anti-human E-Cadherin and PODXL (podocalyxin-like) antibodies.

MG-63 cells were used for screening mouse anti-human GATA1 (GATA binding protein 1) antibody.

Figure 1: Passage of document “16316465” shows annotations for our six entity types. The following colors and abbreviations are used: “anat” or “anatomy” (yellow) for anatomical parts, “cell type” (red) for cell types, “spc” (dark blue) for species, “component” (purple) for cell components, “gene” and “gene or protein” (light blue) for genes and proteins and “c line” (rose) for cell lines. Visualization of the corpus is provided using Brat annotation tool.

named-entity recognition procedures[7] for case-insensitive matching of the synonyms to the text.

Regarding cell lines, we created a dictionary of synonyms by merging names of cell lines from three different sources: hESCReg (Borstlap et al., 2008), a list of human embryonic cell lines presented as supplementary material in (Löser et al., 2010) and data available on-line in the Cell Line Data Base (Romano et al., 2009). Variations for the cell line synonyms were automatically generated (e.g., “CCTL-6”, “CCTL 6”, “CCTL6”). The derived dictionary of synonyms was also matched to the text using Lingpipe.

Besides Metamap, genes were extracted using GNAT (Hakenberg et al., 2008; Hakenberg et al., 2011) configured with the default model. We defined a threshold score of 0.2 for dismissing potential false positives. Finally, mentions for species were extracted using Linnaeus tool (Gerner et al., 2010), besides Metamap.

As discussed before, we have not yet made use of our corpus for training specific taggers for the recognition of any type of entity. Instead, we tried to use freely available systems, terminologies, databases and ontologies. For evaluation of our baseline methods, the corpus was split in two groups, five for development (16316465, 17381551, 17389645, 18162134, 18286199) and five for testing (15971941, 16623949, 16672070, 17288595, 17967047). We have used the development dataset for the error analysis while we kept the other dataset for a blind test.

The evaluation results for each entity type for both datasets are presented in Table 4. Results refer only to the recognition of the mentions, with no normalization of entities. We present results for exact matching, i.e., the exact mention as well as the exact type of entity, and for a more flexible strategy in which we consider also as a correct match any overlapping mention belonging to the same type.

4. Discussion

In this work we have presented a first version of our ongoing corpus developed for the CellFinder project. We have annotated more than 5,200 annotations for a corpus of more than 65,000 tokens, which makes the density of our corpus about 8%. This is a satisfactory density provided that the named entities were usually annotated in the context of biological events. Additionally, no text mining has been performed in the documents before its manual annotations, i.e., the annotators have worked over texts free of any pre-annotations. Finally, although Brat provides a way of querying on-line resources during the annotation (e.g., EntrezGene or Uniprot), it does not support for terminologies and ontologies, which might increase the density of annotations as well as the agreement among annotators.

The density of our corpus is comparable to other full text corpora. For instance, the 14 full papers belonging to the Genia Event Task from the BioNLP Shared Task 2011 also

Entities	Semantic groups
Anatomy	“Anatomical Structure“, ”Body Location or Region“, “Body Part, Organ, or Organ Component”, “Body Space or Junction”, “Body Substance”, “Body System”, “Embryonic Structure”, “Fully Formed Anatomical Structure”, “Tissue”
Cell Comp.	“Cell Component”, “Nucleic Acid, Nucleoside, or Nucleotide”
Cell Type	“Cell”
Gene/protein	“Amino Acid, Peptide, or Protein”, “Enzyme”, “Receptor”, “Amino Acid Sequence”, “Carbohydrate Sequence”, “Gene or Genome”, “Molecular Sequence”, “Nucleic Acid, Nucleoside, or Nucleotide”, “Nucleotide Sequence”
Species	“Amphibian”, “Animal”, “Archaeon”, “Bacterium”, “Bird”, “Eukaryote”, “Family Group”, “Fish”, “Fungus”, “Group”, “Human”, “Mammal”, “Organism”, “Plant”, “Population Group”, “Reptile”, “Vertebrate”, “Virus”

Table 3: Semantic types which have been considered for each entity when using Metamap.

[7] http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html

Entities	Development				Test			
	Exact matches		Overlapping matches		Exact matches		Overlapping matches	
	Recall	F-score	Recall	F-score	Recall	F-score	Recall	F-score
Anatomy	0.30	0.23	0.48	0.33	0.32	0.29	0.48	0.41
Cell Comp.	0.55	0.31	0.67	0.38	0.75	0.17	0.80	0.19
Cell Line	0.48	0.35	0.48	0.35	0.43	0.43	0.62	0.59
Cell Type	0.46	0.42	0.66	0.59	0.37	0.36	0.57	0.55
Gene/Protein	0.68	0.36	0.78	0.44	0.77	0.29	0.90	0.35
Species	0.90	0.46	0.93	0.48	0.83	0.47	0.86	0.49
TOTAL	0.56	0.31	0.69	0.38	0.50	0.28	0.67	0.37

Table 4: Recall and F-score of each entity type for the development and testing datasets. We present results when evaluating using a exact matching and when allowing overlapping of the annotations.

hold a density of about 8% for the proteins. A corpus on microbial cellular responses reported in the work of (Carreira et al., 2011) contains 130 full text documents annotated with 59,000 annotations of biomedical concepts. Although the number of tokens has not to be provided in the publication, its density seems to be lower that 10%. Finally, in the CRAFT corpus (Cohen et al., 2011), which comprise 97 full text documents, annotations have been performed using available ontologies, such as NCBI Taxonomy or Gene Ontology. It contains 597,000 tokens and 118,783 annotations, i.e, a density of almost 20%.

Regarding the disagreement between our annotators, 80% to 85% is also considered satisfactory for the biomedical domain. However, the distinction among anatomical parts, cell components and cell types still need to be discussed further in the next phase of our project. The overlapping among these entities is certainly due to the granularity of our annotation schema, specially on the anatomical level. By manually checking some of the annotations which were only performed by one of the annotators, we have noticed that they do not usually take part on the biological events, which is the final aim of our ongoing corpus. When compared to the CRAFT corpus (Bada et al., 2010), their inter-annotator agreements ranges from 70% to almost 100%, provided that they have performed various training sessions and that the annotation was supported by available ontologies. Our inter-annotator agreement is also comparable to the microbial cellular response corpus (Carreira et al., 2011) which ranges from 21% to 83% after three training cycles.

However, our corpus has some limitations. As already discussed, we have not used any available terminology or ontology while performing the annotations. Therefore, we only provide text mentions, without any association to an identifier.

Being an ongoing project, an extension of the corpus is already being carried out. We have started the annotation of biological events relevant to the stem cell research, such as cell differentiation and gene expression in cells and in anatomical parts. These are valuable information which we plan to make available to the scientific community in our CellFinder project's database, along with the respective bibliographic reference.

We are also proceeding to the annotation of meta-knowledge according to the work of (Thompson et al., 2011). This information is of great importance regarding the reliability of the data being extracted, whether it describes the existence or not of a certain biological process, its intensity (high or low) and the primary publication for finding further information, which is essential when associating data in CellFinder to its respective publication.

Finally, we also intend to annotate a larger number of abstracts in order to have more diversity of entities and biological events. A larger corpus is also usually necessary for training and evaluating machine learning methods for extracting entities or biological events. Our preliminary results show that training a classifier might be necessary at least for the extraction of cell lines (as discussed below).

Regarding the methods and the evaluation presented here, we performed a brief analysis of the errors for all six entities. This analysis was performed only on the five documents belonging to the development corpus. As we are more concerned about the recall of the system, we focused our error analysis on the false negatives. A discussion of the mistakes is presented below for each entity type.

When performing an extra evaluation and allowing overlapping mentions between different types of entities, the recall for the anatomical parts increases from 48% to 65% (result not shown) for the development dataset. Most of these new matches are with annotations which have been extracted by Metamap as cell types, such as "neural" or "myotubes". However, using only Metamap seems to be not enough for achieving a satisfactory recall. Alternative tools, which we plan to use in the next phase, include the recent work of (Pyysalo et al., 2011) on the recognition of anatomical entities using open biomedical ontologies. We are also aware that we cannot expect a high recall from the existing available tools for those entity types which still have a low agreement among the annotators, such as anatomical parts, cell components and cell types.

Our recall is also not enough for the recognition of cell types, whose extraction is based on Metamap and the Cell Type ontology. However, 64% of our false negatives correspond only to the plural forms of common abbreviations in the stem cell types, such as "hNSC" (human neural stem cells) and "hESCs" (human embryonic stem cells). The use of abbreviation resolution methods (Schwartz and Hearst, 2003) in the next phase of the project may help to overcome this problem.

However, Metamap returns a high recall when used for extracting cell components. We consider the recall of 67%-80% as satisfactory for an entity type which plays a secondary role in our annotation schema, as it is not usually associated to a biological event. Additionally, the mapping of the semantic type "Nucleic Acid, Nucleoside, or Nucleotide" to this entity type has increased its recall from 44% (result not shown) to 67% for the development dataset, due to the recognition of annotations such as "DNA", "cDNA" and "mRNA".

Likewise, Linnaeus and Metamap perform very well when extracting species, as they provide a recall of 93% and 86% for the development and test datasets, respectively. Although Linnaeus might be enough for retrieving species, we also consider Metamap because it increased the recall for the test dataset from 82% (result not shown) to 86%. The mentions that are missed are mostly due to problems in the parsing of tables, when columns are concatenated into a single token, such as "hPODXLYesNoNoMouse".

On the other hand, the recall for one of the most important entities in our annotation schema, the cell lines, is still rather low, and about half of the annotations are missed. The two more frequent false negative mentions are "SD56" and "NTERA-2". The first one is not present in any of the three dictionaries. Regarding the "NTERA-2", other cell lines related to it could be found in one of our dictionaries (Cell Line Database) as "NTERA-2 clone D1", but it could not be matched using just a case-insensitive matching strategy. For the cell lines, our baseline approach, which considers only freely available dictionaries and tools, does not seem to address the diversity of the nomenclature. The lack of an integrated cell line database frustrates the hopes of having a more complete terminology of cell line names. The use of a machine learning algorithm trained with some of our annotated documents seems to be inevitable in next phases of the project.

Finally, regarding the extraction of genes and proteins, our recall still need to be improved, as 78% (development dataset) might not be enough for an entity which directly participates in many biological events. On the other hand, GNAT and Metamap have achieved the highest recall for all entities for the test dataset. Surprisingly, Metamap increased the recall from 75% (results not shown) to 90% for the test dataset. By analyzing the false negatives, most of them are never found by GNAT (e.g., "eMyHC" and "TuJ1"). However, some mentions have been missed due to the same problem experienced by Linnaeus, i.e., due to the parsing of the tables, such as the token "hSOX17NoYesYesMs" which contains the gene "SOX17". We plan to try some additional available tools for the extraction and normalization of genes and entities, such as ABNER (Settles, 2005), BANNER (Leaman and Gonzalez, 2008), GeneTuKit (Huang et al., 2011), as well as other resources discussed in (Kabiljo et al., 2009).

In our curation process for the CellFinder project, data extracted using text mining methods will be validated by experts before being included into the database. Therefore, we expect false negatives to be curated manually and false positives to be dismissed or corrected by the curators. Nevertheless, a more precise and high-recall text mining approach will certainly reduce the human effort in the validation step.

Regarding the limitations of our methods and evaluation, as discussed before, our corpus only provides the textual mentions for the annotations. Therefore, we did not consider the normalization of the entities in this phase of the project. However, when mapping data extracted using text mining methods to CellFinder's database, which is completely based on ontologies, the availability of an identifier associated to each entity will become an important issue.

5. Conclusion

In this work we have presented the first version of the corpus which has been annotated in the scope of the CellFinder project. This is an ongoing work which aims to annotate biological processes relevant to the stem cell research. This first version of the corpus includes annotations for six types of semantical entities: anatomical parts (e.g., tissues and organs), cell components, cell lines, cell types, genes/proteins and species. This corpus is composed of 10 full papers which contain around 65,000 tokens and more than 5,200 annotations with an inter-annotator agreement around 80% to 85%. We hope it can be a valuable resource for the stem cell research as well as for evaluation of named-entity recognition methods for a variety of entities.

We have also presented here our baseline methods for the prediction of the entities present in the corpus. We have used only freely available systems, terminologies and ontologies. We have obtained a recall which ranges from 48% to 90%, depending on the entity type. Although some improvements are still necessary regarding the agreement between the annotators and the text mining methods for the prediction of the annotations, the work presented here is promising. We believe that this is a unique corpus in the stem cell domain and the data extracted using literature mining will be a valuable source of information once available in CellFinder's database.

6. Acknowledgements

Mariana Neves was supported by DFG (LE 1428/3-1). We thank Dr. Peter Löser (Robert Koch Institute) and his assistants for the support in the selection of relevant publications for the corpus. We thank Pontus Stenetorp (University of Tokyo) and Sampo Pyysalo (NaCTeM and University of Manchester) for the support with Brat annotation tool and Illés Solt (Budapest University of Technology) for helping in the installation of the tool. We also thank Philippe Thomas and Michael Weidlich, our colleagues in the Humboldt-Universität zu Berlin, for their helpful feedback and support with text mining tools. Finally, we want to acknowledge the other members of the CellFinder project (Jean-Fred Fontaine, Fritz Lekschas, Harald Stachelscheid and Stefanie Werner) for their valuable contribution to this publication.

7. References

Alan R Aronson and Franois-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Anthony Atala. 2012. Regenerative medicine strategies. *Journal of Pediatric Surgery*, 47(1):17 – 28.

Michael Bada, Lawrence E. Hunter, Miriam Eckert, and Martha Palmer. 2010. An overview of the craft concept annotation guidelines. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 207–211, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonathan Bard, Seung Rhee, and Michael Ashburner. 2005. An ontology for cell types. *Genome Biology*, 6(2):R21.

Lucio Barile, Claudia Altomare, and Antonio Zaza. 2011. Induced pluripotent stem cells: progress towards a biomedical application. *Expert Review of Cardiovascular Therapy*, 9(10):1265–1269.

Joeri Borstlap, Glyn Stacey, Andreas Kurtz, Anja Elstner, Alexander Damaschun, Begoa Arn, and Anna Veiga. 2008. First evaluation of the european hescreg. *Nature Biotechnology*, 26:859 – 860.

Rafael Carreira, Sonia Carneiro, Rui Pereira, Miguel Rocha, Isabel Rocha, Eugenio Ferreira, and Analia Lourenco. 2011. Semantic annotation of biological concepts interplaying microbial cellular responses. *BMC Bioinformatics*, 12(1):460.

K Bretonnel Cohen, Helen Johnson, Karin Verspoor, Christophe Roeder, and Lawrence Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

K. Bretonnel Cohen, Tom Christiansen, William A. Baumgartner, Jr., Karin Verspoor, and Lawrence E. Hunter. 2011. Fast and simple semantic class assignment for biomedical text. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laura Furlong, Holger Dach, Martin Hofmann-Apitius, and Ferran Sanz. 2008. Osirisv1.2: A named entity recognition system for sequence variants of genes in biomedical literature. *BMC Bioinformatics*, 9(1):84.

Martin Gerner, Goran Nenadic, and Casey Bergman. 2010. Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85.

Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder, and Graciela Gonzalez. 2008. Inter-species normalization of gene mentions with gnat. *Bioinformatics*, 24(16):i126–i132.

Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M. Bergman. 2011. The gnat library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771.

Akiko Hatano, Hirokazu Chiba, Harry Amri Moesa, Takeaki Taniguchi, Satoshi Nagaie, Koji Yamanegi, Takako Takai-Igarashi, Hiroshi Tanaka, and Wataru Fujibuchi. 2011. Cellpedia: a repository for human cell information for cell studies and differentiation analyses. *Database*, 2011.

Minlie Huang, Jingchen Liu, and Xiaoyan Zhu. 2011. Genetukit: a software for document-level gene normalization. *Bioinformatics*, 27(7):1032–1033.

Marc Jung, Hedi Peterson, Lukas Chavez, Pascal Kahlem, Hans Lehrach, Jaak Vilo, and James Adjaye. 2010. A data integration approach to mapping oct4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells. *PLoS ONE*, 5(5):e10709, 05.

Renata Kabiljo, Andrew Clegg, and Adrian Shepherd. 2009. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10(1):233.

Liz Kerrigan and Raymond W. Nims. 2011. Authentication of human cell-based products: the role of a new consensus standard. *Regenerative Medicine*, 6(2):255–260.

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

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

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.

Roman Klinger, Corinna Kolik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. 2008. Detection of iupac and iupac-like chemical names. *Bioinformatics*, 24(13):i268–i276.

Robert Leaman and Graciela Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium of Biocomputing*, pages 652–663.

Peter Löser, Jacqueline Schirm, Anke Guhr, Anna M. Wobus, and Andreas Kurtz. 2010. Human embryonic stem cell lines and their use in international research. *STEM CELLS*, 28(2):240–246.

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

Sampo Pyysalo, Tomoko Ohta, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Anatomical entity recognition with open biomedical ontologies. In *Proceedings of the Fourth International Symposium on Languages in Biology and Medicine (LBM 2011)*.

Dietrich Rebholz-Schuhmann, Antonio Jimeno-Yepes, Erik M. van Mulligen, Ning Kang, Jan A. Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. Calbc silver standard corpus. *J. Bioinformatics and Computational Biology*, 8(1):163–179.

Paolo Romano, Assunta Manniello, Ottavia Aresu, Massimiliano Armento, Michela Cesaro, and Barbara Parodi. 2009. Cell line data base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Research*, 37(suppl 1):D925–D932.

Barbara Rosario and Marti A. Hearst. 2004. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, pages 451–462.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, Jul.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics. (to appear).

Lorraine Tanabe, Natalie Xie, Lynne Thom, Wayne Matten, and W John Wilbur. 2005. Genetag: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S3.

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

James A. Thomson, Joseph Itskovitz-Eldor, Sander S. Shapiro, Michelle A. Waknitz, Jennifer J. Swiergiel, Vivienne S. Marshall, and Jeffrey M. Jones. 1998. Embryonic stem cell lines derived from human blastocysts. *Science*, 282(5391):1145–1147.

Sowmya Viswanathan and Armand Keating. 2011. Overcoming the challenges of conducting translational research in cell therapy. *Frontiers of Medicine*, 5:333–335. 10.1007/s11684-011-0166-2.

C. J. E. Watson and J. H. Dark. 2012. Organ transplantation: historical perspective and current practice. *British Journal of Anaesthesia*, 108(suppl 1):i29–i42.

Inken Wohlers, Harald Stachelscheid, Joeri Borstlap, Katrin Zeilinger, and Jrg C. Gerlach. 2009. The characterization tool: A knowledge-based stem cell, differentiated cell, and tissue database with a web-based analysis front-end. *Stem Cell Research*, 3(23):88 – 95.

Junying Yu, Maxim A. Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L. Frane, Shulan Tian, Jeff Nie, Gudrun A. Jonsdottir, Victor Ruotti, Ron Stewart, Igor I. Slukvin, and James A. Thomson. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–1920.

Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers

Raheel Nawaz[1], Paul Thompson[1,2], Sophia Ananiadou[1,2]
[1]School of Computer Science, University of Manchester, UK
[2]National Centre for Text Mining, University of Manchester, UK
E-mail: raheel.nawaz@cs.man.ac.uk, paul.thompson@manchester.ac.uk, sophia.ananiadou@manchester.ac.uk

Abstract

Biomedical literature contains rich information about events of biological relevance. Event corpora, containing classified, structured representations of important facts and findings contained within text, provide an important resource for the training of domain-specific information extraction (IE) systems. Such corpora pay little attention to the interpretation of events, e.g., whether an event describes a fact or an analysis of results, whether there is any speculation surrounding the event, etc. These types of information are collectively referred to as *meta-knowledge*. As previous work, an annotation scheme to enrich event corpora with meta-knowledge was designed to facilitate the training of more sophisticated IE systems, and was applied to the complete GENIA Event corpus of biomedical abstracts. In this paper, we describe a case study in which four full papers annotated with GENIA events have been manually enriched with meta-knowledge annotation. We analyse the annotation results, and compare them with the previously annotated abstracts.

Keywords: meta-knowledge, annotation, events, information extraction, biomedical literature

1. Introduction

Due to the rapid growth in the body of scientific literature, it is becoming increasingly important to move beyond simple keyword-based searching to more sophisticated methods that can help researchers to isolate information of interest from a potential mountain of relevant documents. Accordingly, text mining has been receiving increasing interest within the biomedical field (Zweigenbaum et al., 2007). In particular, information extraction (IE) systems produce structured, template-like representations of important facts and findings within documents, called *events*. The extracted events can form the basis of sophisticated semantic search systems, in which users specify search criteria through the (partial) completion of a structured template, which is matched against the extracted events.

IE systems are sensitive to the features of the text on which they operate, and relevant event types vary between domains. Accordingly, such systems must be adapted to deal with specific domains. The usual method of adaptation is the application of machine-learning methods to annotated corpora, e.g. (Soderland, 1999; Califf & Mooney, 2003). In the biomedical field, several corpora annotated with events have been produced, most notably the GENIA event corpus (Kim et al., 2008), the BioInfer corpus (Pyysalo et al., 2007) and the GREC corpus (Thompson et al., 2009). Research into event extraction systems was greatly boosted by the BioNLP'09 shared task on event extraction, in which 24 teams participated (Kim et al., 2009).

Until recently, most event corpora, and thus the systems trained on them, dealt exclusively with abstracts from small subdomains of molecular biology. However, the development of systems that automatically analyse full papers is also vital, given that less than the 8% of scientific claims occur in abstracts (Blake, 2010). However, since there are significant structural and linguistic differences between full papers and abstracts (Cohen et al., 2010), adapting text mining technology from abstracts to full papers presents significant challenges. In terms of event extraction, an effort to move beyond the previous constraints is described in Pyysalo et al. (2010), which concerned the extraction of events from full papers in a new domain, i.e. infectious diseases. This theme was continued in the BioNLP Shared Task 2011 (Kim et al., 2011a), which included tasks relating to four different domains. The original corpus from the BioNLP'09 shared task (derived from the GENIA event corpus) was extended with a small number of full papers annotated according to the same event scheme, to allow evaluation of event extraction technology on full papers (Kim et al., 2011b).

The focus of the annotation in most event corpora is on locating appropriate events in texts, assigning types to them and identifying event participants. However, detailed information about how the events are to be interpreted according to their textual context is usually missing from the annotations. Such information is termed as "meta-knowledge" (Nawaz et al., 2010). Very basic meta-knowledge information is included in most existing corpora, e.g., negated events are identified in BioInfer corpus, whilst negation and basic speculation information are present in the GENIA corpus and the two related corpora from the two BioNLP shared tasks. Such basic meta-knowledge is, however, not sufficient to distinguish between events that express the following types of meta-knowledge:

- Accepted facts vs. experimental findings.
- Hypotheses vs. interpretations of experimental results.
- Previously reported findings vs. new findings.

Previously, an annotation scheme tailored enriching biomedical event corpora with detailed meta-knowledge along five different dimensions was defined (Nawaz et al., 2010). A slightly modified version of the meta-knowledge scheme was subsequently applied to the GENIA Event corpus (1000 MEDLINE abstracts, containing 36,858 events) (Thompson et al., 2011).

In line with the extension of event extraction systems to deal with full papers, it is important to ensure that meta-knowledge can also be assigned to events in full texts. As a first step, we have performed a case study in which we have applied our meta-knowledge scheme to 4 event-annotated full papers. In this paper, we analyse the outcomes of this new meta-knowledge annotation effort, and compare the results to those obtained for abstracts in the GENIA event corpus. It is our intention that insights gained will help to feed into the design of systems that can automatically assign meta-knowledge at the level of full papers as well as abstracts.

2. Event-Based Text Mining

The process of event annotation normally consists of the identification of an event trigger and event participants, and the assignment of types/categories to each of these. The *event-trigger* is a word or phrase in the sentence that indicates the occurrence of the event (often a verb or nominalisation). The *event-type* (generally assigned from an ontology) categorises the type of information expressed by the event. The event participants, i.e., entities or other events that contribute towards the description of the event, are often categorised using semantic role labels such as *cause* and *theme*. Usually, semantic types (e.g. *gene*, *protein*, etc.) are also assigned to the named entities (NEs) participating in the event.

In order to illustrate this typical event representation, consider the following sentence from GENIA Event corpus (PMID: 3035558):

> *The results suggest that the narL gene product activates the nitrate reductase operon.*

Figure 1 shows the typical structured representation of the biomedical event described in this sentence.

TRIGGER: ***activates***
TYPE: **positive_regulation**
THEME: ***nitrate reductase operon:*** **operon**
CAUSE: ***narL gene product*****: protein**

Figure 1: Typical representation of a bio-event

The automatic recognition of such events allows users to create structured queries, on which different kinds of restrictions can be specified to restrict the types of events to be retrieved (Miyao et al., 2006). These restrictions may concern the type of event to be retrieved, the types of participants that should be present in the event or the values of these participants, in terms of either specific strings or NE types.

3. Meta-Knowledge Annotation Scheme

Our event-based meta-knowledge scheme aims to capture as much useful information as possible about individual events from their textual context, to support the training of enhanced event-based search systems. Such enhanced systems could improve the efficiency of tasks such as building and updating models of biological processes, e.g., pathways (Oda et al., 2008) and curation of biological databases (Ashburner et al., 2000; Yeh et al., 2003). Central to both of these tasks is the identification of new knowledge, i.e. experimental findings or conclusions that relate to the current study, and which are stated with a high degree of confidence. Meta-knowledge identification is also useful when checking for inconsistencies or contradictions in the literature, since the meta-knowledge values assigned to two otherwise identical events can affect their interpretation in both subtle and significant ways.

The scheme consists of multiple annotation dimensions to capture different aspects of meta-knowledge. For each dimension, a single category is assigned from a fixed set of possible values. If the category of a given dimension is assigned based on the presence of a particular word or phrase in the sentence, this is also annotated as a “clue”. The scheme was inspired by previous multi-dimensional efforts to assign meta-knowledge to continuous text spans, e.g. (Wilbur et al., 2006; Liakata et al., 2010). The feasibility of automating annotation according to both of these schemes has subsequently been demonstrated (Shatkay et al., 2008; Liakata et al., 2012).

In contrast to the two schemes mentioned above, which concern the annotation of continuous text spans, our meta-knowledge annotation scheme (Thompson et al, 2011) is the first that is specifically tailored to the enrichment of event annotations. In addition to allowing several distinct types of information to be encoded about events, the multi-dimensional nature of the scheme allows the interplay between the different dimension values to be used to derive further useful information (*hyper-dimensions*) regarding the interpretation of the event. The scheme is summarized in Figure 2. A brief overview of the dimensions of our scheme and their possible values are provided below. Each dimension has a default value that is assigned if the event’s textual context does not provide evidence for the assignment of one of the other values.

Knowledge Type (KT): Captures the general information content of the event. Each event is classified as one of the following: *Investigation* (enquiries and examinations), *Observation* (direct experimental observations), *Analysis* (inferences, interpretations and conjectures), *Method* (experimental methods) *Fact* (general facts and well-established knowledge) or *Other* (default: events expressing incomplete information, or whose KT is unclear from the context)

Certainty Level (CL): Encodes the confidence or certainty level ascribed to the event in the given text. We partition the epistemic scale into three distinct levels: *L3* (default: no expression of uncertainty), *L2* (high confidence or slight speculation) and *L1* (low confidence or considerable speculation).

Polarity: Identifies negated events. We define negation as the absence or non-existence of an entity or a process. Possible values are *Positive* (default) and *Negative.*

Manner: Captures information about the rate, level, strength or intensity of the event, using three values: *High* (the event occurs at a high rate or level of intensity), *Low* (the event occurs at a low rate or level of intensity) or *Neutral* (default: no indication of rate/intensity).

Source: Encodes the source of the knowledge being expressed by the event as *Current* (default: the current study) or *Other* (any other source).

Hyper-Dimensions: Correspond to additional information that can be interfered by considering combinations of some of the explicitly annotated dimensions. We have identified two such hyper-dimensions each with binary values (*Yes* or *No*): *New Knowledge* (inferred from *KT*, *Source* and *CL*) and *Hypothesis* (inferred from *KT* and *CL*).

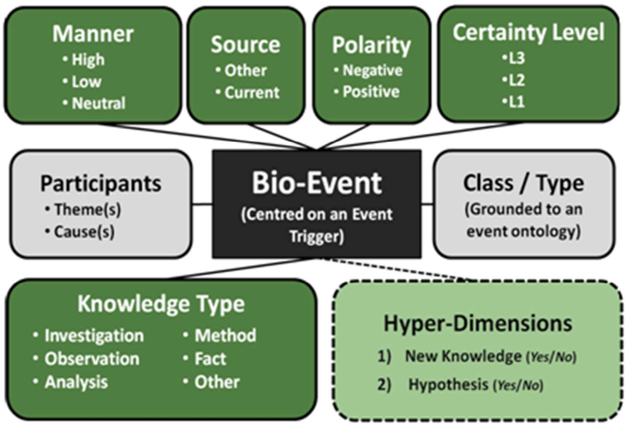


Figure 2: Meta-knowledge annotation scheme

The annotation of the GENIA Event corpus according to this scheme (Thompson et al., 2011) showed that high levels of inter-annotator agreement (between 0.843 and 0.929 Kappa) were achieved by following the 66-page guidelines. Also, given that each of the two annotators had a different background (biology vs. linguistics), it was concluded that specific expertise does not appear necessary to perform meta-knowledge annotation.

In the context of the current case study, it was important to consider whether the meta-knowledge scheme needed to be altered prior to its application to full papers. This consideration is relevant, firstly due to the fact that the scheme was defined only on the basis of examining abstracts, and secondly since previous research into meta-knowledge classification at the sentence or zone level has defined different numbers and types of categories to encode the general information content of the sentence/zone, according to whether abstracts (e.g. (McKnight & Srinivasan, 2003; Ruch et al., 2007; Hirohata et al., 2008)) or full papers (e.g. (Mizuta et al., 2006; Liakata et al., 2010)) are under consideration. For full papers, the number of categories defined can be more than double the number used for abstracts.

The information encoded by the *KT* dimension of the event-based meta-knowledge scheme is somewhat comparable to the above schemes. However, while sentence-based categories are quite strongly tied to structural aspects of the article, with labels such as *background, experiment, conclusion,* etc., the values of the *KT* dimension can be considered more abstract or high level. For example, if several different events occur in *background* and *conclusion* sentences, each event could be assigned a different KT value. That is to say, both sentence types could contain certain events that describe observations, and others that represent analyses. Due to the more abstract level of information encoded by *KT* types, we believe them to be applicable both to abstracts and full papers. They can be considered as complementary to sentence or zone-based schemes, in allowing a finer-grained analysis of the different types of information that can occur within a particular sentence or zone type.

We also envisage that the other dimensions of the scheme do not need to be expanded to allow annotation of full papers, as they all appear to represent general features that can be found in many types of text. For example, the use of three different levels of certainty is in line with an analysis of general characteristics of the English language (Hoye, 1997), rather than being specific to abstracts. The two-way distinctions of the *Polarity* and *Source* dimensions are also observable in any kind of academic writing. Similarly, the information encoded by the *Manner* dimension, whilst more domain specific, should also be applicable to full papers.

The ability to apply the same meta-knowledge scheme to both abstracts and full papers has advantages not only in terms of comparing meta-knowledge characteristics between the two text types, but also in facilitating easy portability/scalability of systems trained to assign meta-knowledge to events either at the abstract or full paper level. In performing meta-knowledge annotation of full papers, careful consideration was given as to whether any aspects of event interpretation were missing from the scheme, or whether there were any events that could not be correctly characterised by the existing categories within the dimensions.

4. Annotation of Full Papers

We have applied our meta-knowledge annotation scheme to four full papers, which had previously been manually annotated with events, according to the GENIA event annotation scheme (Kim et al., 2008). According to the previously proven consistency of the meta-knowledge annotation that can be achieved by following the guidelines (Thompson et al., 2011), regardless of annotator background, the meta-knowledge annotation was carried out manually by one of the authors, who has a background in computational linguistics. All events in the four papers were annotated with meta-knowledge, without any concerns regarding deficiencies in the existing scheme, either in terms of missing dimensions, or missing values in existing dimensions. This suggests that the scheme is fully portable between abstracts and full papers.

Table 1 summarises the distribution of the annotations amongst the different categories for each dimension, and Table 2 shows the most frequent clues for each category and their relative frequencies, i.e., the percentage of events of the specified category in which the clue is annotated. Below, we provide a brief discussion of the results of our new annotation effort. We examine results at the level of the complete papers, and also consider the distributions of annotations within the major sections of the papers, i.e., *Background, Methods, Results, Discussion* and *Conclusion.*

4.1 Knowledge Type (KT)

The most commonly annotated value is *Observation*, constituting just over a third of the total number of events. This is unsurprising, since a large proportion of most biomedical papers would be expected to report on definite experimental observations and results.

Considering individual sections within the full papers, *Observation* events are most prevalent in *Background* (42% of all events in this section type). It may seem surprising that the frequency of *Observation* events in *Background* is greater than in *Results.* However, *Observation* events can refer to previous work as well as current work, and the *Background* section will often refer to findings from a large number of related studies. In the *Results* section, approximately 36% of events describe observations; while in the *Discussion* section, the frequency of such events is even lower (32%). This is to be expected, since greater proportion of this section type would normally be analytical in nature.

Only in a small fraction (12%) of the *Observation* events is the *KT* type determined by the presence of an explicit lexical clue (mostly sensory verbs). In most cases, the tense of the event-trigger and the context of the event (both local and global position within the paper) were found to be important factors.

The second most prevalent category is *Other*. These events generally constitute participants of other events whose *KT* value is *Investigation*, *Analysis* or *Fact*. Out of the context of their parent event, these participant events have no specific *KT* interpretation. No explicit lexical clues were annotated for this category.

A relatively large proportion of events (more than one fifth) belong to the *Analysis* category. This makes sense, given that analytical elements are normally to be found to some extent in most section types in full papers. These include the *Background* section, where such events are most likely to provide overviews or interpretations of previous work, as well the *Results*, *Discussion* and *Conclusions* sections, where analyses, interpretations and conclusions regarding authors' own work most commonly appear. As may be expected, the frequency of *Analysis* events is highest in *Discussion/Conclusion* sections, where they constitute over one quarter (27%) of all events.

An explicit lexical clue was found for each *Analysis* event. The clues comprised verbs, modal auxiliaries and certain adverbs (such as, *thus* and *therefore*).

Almost 6% of the events belong to the *Method* category. Although full papers generally include a fairly large *Methods* section, the small number of events falling into this category is largely because the GENIA event annotation focusses on dynamic relations, i.e., at least one of the biological entities in the relationship is affected, with respect to its properties or its location, in the reported context. This means that descriptions of methods are often less relevant event annotation targets than are events describing observations and analyses.

Our case study suggests that only a small proportion of events in full papers (around 4%) describe factual knowledge. Such events are not evenly distributed throughout papers, and occur most frequently in *Background* (7.5% of all events in this section type), in order to provide context for the new research described in the paper. They can also appear in the *Discussion* section (4.5% of events), where they may be contrasted or compared with the outcomes of the current study. As may be expected, factual knowledge is almost never referred to in the *Results* sections of papers. Similarly to the *Observation* category, most (85%) events from this category did not have an explicit lexical clue.

Dimension	Category	Events	Relative Frequency (RF)
Knowledge Type (KT)	Analysis	381	22.3%
	Investigation	65	3.8%
	Observation	619	36.2%
	Fact	70	4.1%
	Method	100	5.8%
	Other	475	27.8%
Certainty Level (CL)	L1	39	2.3%
	L2	162	9.5%
	L3	1509	88.2%
Polarity	Negative	63	3.7%
	Positive	1647	96.3%
Manner	High	66	3.9%
	Low	15	0.9%
	Neutral	1629	95.3%
Source	Current	1369	80.1%
	Other	341	19.9%
Hyper-Dimensions	New Knowledge	489	28.6%
	Hypothesis	259	15.1%

Table 1: Category distribution

The *Investigation* KT category is the least frequent. The results of our annotation experiment suggest that the *Background* section normally very briefly introduces the subject of investigation (2.5% of events in this section type). A slightly more detailed description of the investigation is then given in the *Results* section (5.4% of all events in this section type). It is also possible that the research goal will be very briefly reintroduced in the

Discussion section of the paper (an average of 1.8% of all events in this section type). All *Investigation* events were accompanied by an explicit lexical clue.

4.2 Certainty Level (CL)

Almost 12% of all events in our full paper sample are expressed with some degree of uncertainty, almost all of which belong to the *KT* type *Analysis*. Taking this into account, the need for this dimension becomes more apparent: whilst under half of *Analysis* events (47%) are stated with no uncertainty, this also means that over a half of these events do express some kind of uncertainty. In fact, 43% of all *Analysis* events are annotated as having slight speculation (*L2*), whilst 10% are reported with greater speculation (*L1*). The marking of uncertainty is sometimes necessary in scientific research literature. Analyses of experimental results may constitute important outcomes, but yet the authors are not confident that their analysis is completely reliable. As stated by Hyland (1996), “Scientists gain credibility by stating the strongest claims they can for their evidence, but they also need to insure against overstatement.” (p. 257). Authors often achieve this by using slight hedging (*L2*). Greater speculation (*L1*) is less common, as credibility is reduced in this case.

Considering individual sections helps to confirm Hyland’s statement. Although the proportion of *Analysis* events that are assigned a *CL* value of *L1* is fairly constant in the *Background, Results* and *Discussion* sections, the proportions of *L2* events have more variation. The relative frequency is lowest in the *Background* sections (36% of *Analysis* events). Since this type of section deals mainly with reporting the work of others, there may be less need to hedge, as it is not the authors’ own credibility at stake. In contrast, the relative frequency of slightly hedged *Analysis* events is noticeably higher in the *Results* and *Discussion* sections (46% and 51%), respectively, where the authors’ own work is the main focus, and hence interpretations and analyses of results are often stated more tentatively.

In terms of clues, modal auxiliaries account for most (70%) of the *L1* events, while the clues for *L2* include both verbs and modals.

4.3 Polarity

Just under 4% of all events are negated. Almost all negated events belong to the KT categories of *Observation* or *Analysis*, which is fairly intuitive. One would not, for example, expect to encounter many cases where *Investigation* or *Method* events are negated. The distributions of negated events vary across different sections of the full papers. The proportions encountered in *Background* and *Discussion* sections are quite similar to each other (around 2% in each section), compared to around 6% of negated events in *Results* sections. Thus, it appears that it is very rare for anything other than positive results to be mentioned in the former two section types. In contrast, when reporting directly on one’s own experimental results, negative results are mentioned more frequently.

Although several negation clues were annotated, the adverbial *not* accounts for over half of negated events.

Dimension	Category	Most Frequent Clues and their RF
Knowledge Type	Analysis	show (16%), demonstrate (14%), indicate (9%), suggest (7%), reveal (5%), can (4%), thus (3%), may (3%)
	Investigation	determine (19%), analyze (15%), elucidate (11%), evaluate (9%), detect (5%), indicate (5%), test (5%), examine (3%), investigate (3%)
	Observation	observe (4%), find (3%), show (1%), document (1%), exhibit (1%)
	Fact	known (6%), well established (3%), well known (2%), fact (2%)
Certainty Level	L1	may (54%), can (15%), possibility (10%), not clear (5%), not understood (5%)
	L2	indicate (22%), can (15%), suggest (11%), ability (6%), able (6%), potential (4%), hypothesize (3%), imply (3%), suspect (3%)
Polarity	Negative	not (57%), no (18%), failure (10%), non (8%), fail (2%), inability (2%)
Manner	High	significantly (17%), well (12%), much (11%), n-fold (9%), strong (9%), strongly (6%), high (3%), higher (3%)
	Low	minimal (13%), little (13%), weak (13%), weaker (13%), n% (7%), less (7%)
Source	Other	Citation (78%), has been (12%), previously (2%), recently (2%)

Table 2: Most frequent clues for each category together with relative frequencies (RF)

4.4 Manner

Almost 5% of all events are expressed with a *Manner* other than *Neutral*. This proportion is fairly constant in the *Background, Results* and *Discussion* sections of the full papers, showing that, although fairly rare, information about the manner of events can be of relevance to the discussion in various different parts of the paper. However, the expression of *High* manner is 4 times more frequent than that of *Low* manner. Similarly to negation, most *High* and *Manner* events belong to *KT* categories of *Observation* or *Analysis*.

Another similar pattern to the *Polarity* dimension is that events with a *Manner* value of *Low* seem to appear with any regularity only in the *Results* sections of the papers, where they appear with just over half the frequency of events whose *Manner* value is *High.* In contrast, the *Low* value was never annotated in the *Background* sections of the papers, and was only annotated for less than 1% of events in the *Discussion* sections. This suggests that events with *Low* manner constitute fairly insignificant information, and are normally mentioned only when reporting experimental results.

Most manner clues are adverbs or adjectives; however numerical values (such as, *n-fold* and *n%*) are also used to express *High* manner.

4.5 Source

Nearly 20% of all events in the full papers belong to the *Other* category. The concentration of such events is highest in the *Background* sections of the papers, where over 40% of the events are attributed to other sources. This is expected, since the *Background* section normally contains the highest concentration of descriptions of previous work. The *Discussion* sections of the papers also have a high (over 25%) concentration of *Other* events, since in this type of section, it is common to compare and contrast the outcomes of the current work with those of previous, related studies. The frequency of *Other* events in the remaining sections is considerably lower. For example, in the *Results* sections of the papers, less than 7% of events are annotated as *Other.* While citations accounted for most of the *Other* events, the use of past perfect tense and explicit markers (such as *previously* and *recently*) also served as clues.

4.6 Hyper-Dimensions

Using the annotations for *KT*, *CL* and *Source* dimensions, we computed the values for the *New Knowledge* and *Hypothesis* dimensions. We found that nearly 29% of all events conveyed new knowledge, and over 15% of all events represented hypotheses. Events conveying new knowledge were predominantly found in the *Results*, *Discussion* and *Conclusion* sections, while hypotheses were found in these sections as well as in the *Background* section. The *Methods* section contained hardly any hypotheses or claims of new knowledge.

5. Comparison with Abstracts

In this section, we compare the distribution of meta-knowledge annotation results obtained in our case study of full papers with those obtained for abstracts, as reported in Thompson et al. (2011). Table 3 shows the difference between the category distributions for full papers and abstracts. Below, we provide a brief discussion of the differences in each dimension.

KT: The biggest difference is seen for the *Method* events, which are more than twice as abundant (in terms of relative frequency) in full papers than in abstracts. This is probably because abstracts tend to focus more on results and their significance, rather than how these results were obtained. As mentioned above, however, the frequency of *Method* events is quite low even for full papers, due to the "dynamic" nature of GENIA events.

A further feature of abstracts is that they tend to contain one or two sentences summarising current knowledge (i.e., well known facts) in the relevant field. Since the average size of abstracts in the GENIA event corpus is 9 to 10 sentences (Kim et al., 2008), the relative frequency of facts in abstracts is quite high (over 8%). This proportion is comparable to the number of factual events in *Background* sections of full papers (over 7% of all events in this section type), where the current state of knowledge is also discussed in some detail. However, as was explained in section 4.1, events describing facts are far scarcer in the other sections of full papers and, given the overall length of papers, the relative frequency of *Fact* events in full papers as a whole is only around half of the frequency in abstracts.

Regarding *Investigation* events, their relative frequency in the *Results* sections of the full papers is comparable to their relative frequency in abstracts (around 5%). However, similarly to the *Fact* category, the extremely rare appearance of *Investigation* events in other sections of full papers means that overall relative frequency in full papers is also much lower than in abstracts.

The relative frequency of *Analysis* events is around 25% higher in full papers than in abstracts. As explained in the previous section, and in contrast to *Fact* and *Investigation* events, *Analysis* events are found with quite high frequency in several sections of full papers. For the *Other* and particularly the *Observation* categories, there is much less variation between the relative frequencies in full papers and abstracts. Thus, clear reporting of experimental observations is equally important throughout both full papers and abstracts.

CL: Owing to the very nature of abstracts, a high proportion of events with no uncertainty is to be expected. As authors aim to "sell" the most positive aspects of their work in abstracts, it makes sense that the majority of analyses should be presented in a confident manner. However, as explained in section 4.2, authors tend to be more cautious while detailing their results and findings in the main body of papers, in order to maintain credibility in case their results are later disproved. The fact that the proportion of slightly hedged *Analysis* events is particularly high in the *Results*, *Discussion* and *Conclusion* sections of full papers, rising as high as 51% in the *Discussion* sections, helps to explain why *L2* events are over 57% more frequent in full papers than in abstracts. The relative frequency of *L1* events is also higher in full papers by about 10%.

Polarity: The relative frequency of negated events is significantly (67%) higher in abstracts than in full papers. This is partly due to the fact that negative results are sometimes more significant than positive results (Knight, 2003), and are therefore, highlighted in the abstracts. In addition, since negated events only appear with any regularity in the *Results* sections of full papers, this helps to explain their lower relative frequency than in abstracts when the complete paper is considered.

Manner: The distribution of *High* and *Neutral* manner is very similar in abstracts and full papers, and the distribution of *Low* manner is exactly same. This follows the same trend described in section 4.4, where it was also noted that the proportions of events with explicit manner markings are also fairly similar across several individual section types within full papers.

Dim.	Cat.	RF (FP)	RF (A)	Diff. in RF (FP – A)	% Change in RF
KT	Ana.	22.2%	17.8%	4.4%	24.8%
	Inv.	3.8%	5.3%	-1.5%	-39.0%
	Obs.	36.3%	34.7%	1.4%	4.1%
	Fact	4.1%	8.1%	-4.0%	-98.7%
	Meth.	5.8%	2.6%	3.2%	120.8%
	Oth.	27.8%	31.3%	-3.5%	-12.7%
CL	L1	2.3%	2.1%	0.2%	9.7%
	L2	9.5%	6.0%	3.5%	57.6%
	L3	88.2%	91.9%	-3.7%	-4.2%
Pol.	Neg.	3.6%	6.1%	-2.5%	-66.7%
	Pos.	96.4%	93.9%	2.5%	2.6%
Man.	High	3.9%	3.8%	0.1%	2.2%
	Low	0.8%	0.8%	0.0%	0.0%
	Neut.	95.2%	95.3%	-0.1%	-0.1%
Src	Cur.	80.0%	98.5%	-18.5%	-23.1%
	Oth.	20.0%	1.5%	18.5%	1248.6 %
Hyper -D	N.K	28.6%	43.4%	-14.8%	-51.7%
	Hypo.	15.2%	13.4%	1.8%	13.4%

Table 3: Difference between relative frequencies (RF) of categories in full papers (FP) and abstracts (A)

Source: This is the dimension for which the largest difference in category distribution exists between abstracts and full papers. Full papers contain 12.5 times as many *Other* events as abstracts. This is mainly because abstracts are meant to summarise the work carried out in the current study. Furthermore, citations, which are the most common way to denote previous work, are often not allowed in abstracts. In contrast, full papers normally mention related work quite extensively, most notably in *Background* and *Discussion* section.

Hyper-Dimensions: While the relative frequency of *Hypothesis* events is higher in full papers, the proportion of *New Knowledge* events is significantly higher in abstracts. This is mainly because, in abstracts, authors typically include most of new discoveries and results, while only mentioning the main hypotheses.

6. Conclusion

In this article, we have described a case study to investigate the feasibility of applying an event level meta-knowledge annotation scheme (Thompson et al, 2011), whose design was originally guided only by reference to abstracts, to full papers. This is important, given that work on event extraction is gradually being scaled from abstracts to full papers, and also that the automatic recognition of meta-knowledge about events can be highly useful for building more sophisticated IE systems. Our case study involved the annotation of 4 full papers using the meta-knowledge annotation guidelines described in Thompson et al. (2011). The results of the case study strongly suggest that the existing meta-knowledge annotation scheme can be successfully applied to full papers, without any modifications

In order to help to guide the engineering of features for event-based meta-knowledge assignment systems trained on full papers, we conducted an analysis of the meta-knowledge annotations created during our case study. The analysis was concerned not only with the overall distribution of meta-knowledge categories in the full papers, but also with comparisons of the distributions of meta-knowledge categories, both between different sections of the papers, and also with meta-knowledge annotations added to the GENIA Event corpus of MEDLINE abstracts (Thompson et al., 2011). In certain cases, notable differences in the distribution of categories within particular dimensions could be observed both between the different sections of full papers, as well as between full papers and abstracts. This suggests that it may be appropriate to train separate meta-knowledge classifiers for full papers and abstracts. It may also be advantageous to use section-specific classifiers within full papers.

Based upon the demonstrated applicability of the meta-knowledge annotation scheme to full papers, we plan to embark upon a larger annotation effort to enrich all full papers from the BioNLP 2011 GENIA event task with meta-knowledge annotation, in order to increase the amount of annotated data available for training meta-knowledge assignment systems that can operate on full papers. We will also aim to enrich other event-annotated corpora released as part of other tasks in the BioNLP 2011 Shared Task, which include both full papers and abstracts dealing with different domains.

7. Acknowledgements

The work described in this paper has been funded by the Biotechnology and Biological Sciences Research Council through grant numbers BBS/B/13640, BB/F006039/1 (ONDEX) and through the MetaNet4U project (ICT PSP Programme, Grant Agreement: No 270893).

8. References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1): 25-29.

Blake, C. (2010). Beyond genes, proteins, and abstracts:

Identifying scientific claims from full-text biomedical articles. *J Biomed Inform*, 43(2): 173-189.

Califf, M.E. and Mooney, R.J. (2003). Bottom-up relational learning of pattern matching rules for information extraction. *The Journal of Machine Learning Research*, 4: 177-210.

Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C. and Hunter, L.E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11: 492.

Hirohata, K., Okazaki, N., Ananiadou, S. and Ishizuka, M. (2008). Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 381-388.

Hoye, L. (1997). *Adverbs and modality in English*, Longman.

Hyland, K. (1996). Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication*, 13(2): 251-281.

Kim, J.-D., Ohta, T. and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9: 10.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on BioNLP: Shared Task*, pp. 1-9.

Kim, J.D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N. and Tsujii, J. (2011a). Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 1-6.

Kim, J.D., Wang, Y., Takagi, T. and Yonezawa, A. (2011b). Overview of Genia Event Task in BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 7-15.

Knight, J. (2003). Null and Void. *Nature*, 422: 554-555.

Liakata, M., Teufel, S., Siddharthan, A. and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC 2010*, pp. 2054-2061.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C. and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*, 28(7).

McKnight, L. and Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *AMIA Annu Symp Proc*, pp. 440-4.

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. and Tsujii, J. (2006). Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of ACL*, pp. 1017-1024.

Mizuta, Y., Korhonen, A., Mullen, T. and Collier, N. (2006). Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6): 468-487.

Nawaz, R., Thompson, P., McNaught, J. and Ananiadou, S. (2010). Meta-Knowledge Annotation of Bio-Events. In *Proceedings of LREC*, pp. 2498-2507.

Oda, K., Kim, J.-D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y. and Tsujii, J.i. (2008). New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics*, 9(Suppl 3): S5.

Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J. and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8: 50.

Pyysalo, S., Ohta, T., Cho, H.C., Sullivan, D., Mao, C., Sobral, B., Tsujii, J. and Ananiadou, S. (2010). Towards event extraction from full texts on infectious diseases. In *Proceedings of the BioNLP 2011 Workshop*, pp. 132-140.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D. and Lovis, C. (2007). Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics*, 76(2-3): 195-200.

Shatkay, H., Pan, F., Rzhetsky, A. and Wilbur, W.J. (2008). Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18): 2086-2093.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1): 233-272.

Thompson, P., Iqbal, S.A., McNaught, J. and Ananiadou, S. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, 10: 349.

Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12: 393.

Wilbur, W.J., Rzhetsky, A. and Shatkay, H. (2006). New directions in biomedical text annotations: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7: 356.

Yeh, A.S., Hirschman, L. and Morgan, A.A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, 19(Suppl 1): i331-i339.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. and Cohen, K.B. (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform.*, 8(5): 358-375.

Clinical Corpus Annotation: Challenges and Strategies

Fei Xia[1,2] **and Meliha Yetisgen-Yildiz**[2,1]
[1]Department of Linguistics, [2]Biomedical and Health Informatics
University of Washington, Seattle, WA 98195, USA
E-mail: {fxia,melihay}@uw.edu

Abstract

Annotation is an important task for Natural Language Processing (NLP), and the traditional annotation schema, including writing detailed guidelines and training annotators, has proved to work well in many previous annotation projects. However, making medical judgment on clinical data requires medical expertise and annotation can only be done by experts. Recently, we created three corpora for our clinical NLP studies: one marks critical recommendations in radiology reports, and the other two indicate whether a patient has pneumonia based on chest X-ray reports or ICU reports. All the annotations were done by medical experts. In this paper, we discuss various challenges we have encountered when dealing with expert annotation, and lay out some lessons we have learned from the annotation tasks. Our experiments show that medical training alone is not sufficient for achieving high inter-annotator agreement, and NLP researchers should get involved in the annotation process as early as possible despite their lack of medical training.

Keywords: Clinical corpus annotation, Annotation schemata, Annotation guidelines

1. Introduction

Over the last decade, Electronic Medical Record (EMR) systems have become increasingly integral to the provision of health care services. Accessibility to the details of patient data available in EMR systems is critical to improve the health care process and advance clinical research. However, most patient information that describes patient state, diagnostic procedures, and disease progress is represented in free-text form. Several studies demonstrated the value of Natural Language Processing (NLP) in extracting the knowledge from clinical records for a variety of health care applications including decision support tools, quality improvement initiatives, and automated encoding for clinical research (Chapman and Cohen, 2009).

Although the premise of NLP is to develop automated approaches to process free-text data available in medical records, building those approaches requires a substantial amount of manual analysis and annotation of data. Roberts et al. (2009) summarized the reasons for the need of manual annotation as: (1) creating annotation scheme serves to focus and clarify the information requirements of the text processing task and the domain of interest, (2) annotated data provides a gold standard to assess the performance of the text processing systems, and (3) annotated data serves as a resource for developing rule-based systems or creating statistical models by the application of machine learning approaches. Therefore, producing high quality annotations is essential to building successful text processing systems.

The traditional annotation schema in the NLP field includes detailed annotation guidelines, well-trained annotators, double annotation, and adjudication. However, when the annotated data are clinical records and the annotation requires substantial medical expertise, we face new challenges while trying to follow the traditional annotation schema. In this paper, we discuss three annotation tasks that use clinical data, describe various challenges we encounter, and lay out some lessons we have learned from the tasks.

2. Related Work

In this section, we discuss common annotation schemata and related work on annotating biomedical data.

2.1 Three Annotation Schemata

Annotation is important for NLP research. Traditionally, an annotated corpus is created by a team consisting of guideline designers, annotators, language or domain experts, and technical support staff. Detailed annotation guidelines are created before annotation starts and they are revised during annotation if necessary. The annotators in this team are often trained on the annotation task for a long period of time. We refer to this approach as *traditional annotation* schema. Almost all the large-scale annotated corpora used in the NLP community were created this way, including the Prague Dependency Treebank (Hajic, 1999), the English/Chinese/Arabic Penn Treebank (Marcus et al., 1993; Xia et al., 2000; Maamouri and Bies, 2004), the English PropBank (Palmer et al., 2005), and the Penn Discourse Treebank (Miltsakaki et al., 2004).

One issue with the traditional annotation scheme is the high cost of training and maintaining annotators. Recently, two other annotation schemata are proposed to address this issue. The first one, *crowd-sourced annotation*, takes advantage of online labor markets such as Amazon's Mechanical Turk (AMT). Because the cost of labors from such markets is much lower than that of well-trained annotators, corpus developers can afford to have multiple annotations on the same data and use majority voting to choose gold standard. This schema can produce good results with low cost when the task is relatively simple and does not require much domain knowledge (Snow et al., 2008). AMT has been applied to the biomedical domain successfully for named-entity extraction tasks from clinical trial descriptions (Yetisgen-Yildiz et al., 2010). The second schema, *community annotation*, gathers annotation from a research community; one example is the evaluation corpus used in the 2009 i2b2 medication challenge, which was created by the i2b2 organizers (who created annotation guidelines and some initial annotations) and the participating teams. This schema can produce good annotation fast and with low cost, if the annotation is careful coordinated and receives strong support from the community (Uzuner et al., 2010).

Corpus	Report Type	Corpus size	Annotation	Annotation Unit	Annotators
C1	Radiology reports	800 reports	critical recommendation	sentence	one radiologist, one internal medicine physician
C2	Chest x-ray reports	1344 reports	PNA and CPIS	report	one general surgeon, one data analyst
C3	Eight ICU report types	5313 reports for 426 patients	PNA	patient	one research study nurse

Table 1. The three corpora in our study. “PNA” stands for “pneumonia”, “CPIS” stands for “Critical Pulmonary Infection Score”, and “ICU” stands for “intensive care unit”.

2.2 Annotating Biomedical Data

Data in the biomedical domain can be divided into the two types. The first type is the collection of research articles in the biomedical literature such as Medline. The second type is clinical patient data such as radiology reports. For the first type, there are various corpora generated from the research articles available in Medline for information extraction tasks on biological events, entities and their interactions. Some well-known, publicly available biomedical corpora include GENIA corpus (Kim et al., 2003), PennBioIE corpus (Kulick et al., 2004), Yapex corpus (Franzen et al., 2002), and GENETAG (Tanabe et al., 2005).

For clinical data, the number of publicly available annotated corpora is quite limited due to concerns regarding patient privacy as well as concerns about revealing unfavourable institutional practices (Chapman et al., 2011). The i2b2 NLP challenges contribute to the clinical NLP research by releasing corpora composed of de-identified clinical records annotated for various different information extraction tasks including smoking history extraction (Uzuner et al., 2008), comorbidity extraction (Uzuner, 2009), named-entity extraction (medication, treatment, test, medical condition) (Uzuner et al., 2010; Uzuner et al., 2011), assertion and relation extraction (Uzuner et al., 2011).

There are other studies on annotating clinical data that are not publicly available. While some of those clinical corpora are about traditional NLP annotations such as POS tagging (Pakhomov et al., 2006) and anaphoric relations (Savova, 2011), other corpora require annotators to be medical experts. These annotation tasks are more domain specific, focusing on the annotation of medical knowledge in clinical text. One example is the Clinical E-Science Framework (CLEF) corpus (Roberts et al., 2007; Roberts et al., 2009). The purpose of the CLEF project is to build a framework for the capture, integration and presentation of clinical information to be used in clinical research, evidence-based health care, and genotype-phenotype mapping. The corpus includes various types of clinical records annotated for named entities and their relations, modifiers, and co-references. Because of the nature of the clinical research, most of the corpora generated in this domain are very specific to a disease or a disease type. For example, Fiszman et al. (2000) annotated chest x-ray reports for automatic identification of acute bacterial pneumonia; South et al. (2009) manually annotated clinical records to identify phenotypic information for inflammatory bowel disease. Fiszman et al.’s annotation was at the report level, whereas South et al.’s annotation was at the phrase level. These three corpora require significant medical knowledge, and the corpora we build in our research projects fall into this category.

3. Our Projects and Corpora

In this section, we discuss three projects that we are currently working on. For each project, we created a corpus to train and evaluate our NLP systems. All three projects deal with patient medical reports, and the corpora were annotated by physicians. The retrospective review of the reports in the corpora was approved by the Human Subjects Committee of Institutional Review Board (IRB) at our institute, who waived the need for informed consent. Table 1 provides a summary of the corpora. In the rest of the section, we will provide a background of the projects, a description of the corpora, and some preliminary results of our NLP systems.

3.1 Critical Recommendations in Radiology Reports

Radiology reports include the descriptions of relevant disease processes found by radiologists on imaging studies, such as radiographs and computed tomography (CT) scans. If a radiologist makes a potentially important observation when examining an imaging study, he/she may include in his/her report further specific recommendations for follow-up imaging tests, or clinical follow-up. These recommendations are made when the radiologist considers the finding to be clinically significant and unexpected, and believes that it is important for the referring physician to consider further investigation, management, or follow-up of the finding in order to avoid an adverse outcome. The American College of Radiology (ACR) recommends that radiologists supplement their written report with “non-routine” means of communication with the referring physician (usually verbal) to ensure adequate receipt of the critical information in a timely manner[1]. Despite the imperative of good communication to avoid medical errors, it does not always occur. Inadequate communication of critical results is the cause of the majority of malpractice cases involving radiologists in the USA (Towbin et al., 2011). The Joint Commission reported that up to 70% of sentinel medical errors were caused by communication errors (Lucey and Kushner, 2010).

[1] ACR practice guideline for communication of diagnostic imaging findings. Available at: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/guidelines/dx/comm_diag_rad.aspx

The goal of our first project is to build an NLP system that automatically identifies critical recommendations in radiology reports so that these recommendations will be highlighted to reduce the chance that they are overlooked by the referring physicians. We defined *critical recommendation* as a statement made by the radiologist in a radiology report to advise the referring clinician to further evaluate an imaging finding by either other tests or further imaging. An example sentence annotated as critical recommendation from our corpus is "*Recommend non-emergent pelvic ultrasound for further evaluation to exclude cystic ovarian neoplasm*."

In order to train and evaluate our system, we created a corpus of radiology reports composed of 800 de-identified radiology reports extracted from Harborview Medical Center radiology information system. Two annotators, one radiologist and one clinician, went through each of the 800 reports and marked the sentences that contained critical recommendations. Out of 18,748 sentences in the reports, the radiologist annotated 118 sentences and the clinician annotated 114 sentences as recommendation. They agreed on 113 of the sentences annotated as recommendation.

Using the corpus, we built a statistical text processing system to classify each sentence in radiology reports as either containing or not containing critical recommendation. The system achieved 95.60% precision, 79.82% recall, and 87% F-score (5-fold cross validation) in identifying recommendation sentences. More detail of the system design and evaluation was reported in (Yetisgen-Yildiz et al., 2011a).

3.2 PNA and CPIS in Chest X-ray Reports

Early detection and treatment of ventilator associated pneumonia (VAP), the most common healthcare associated infections in critically ill patients, is important; even short-term delays in appropriate antibiotic therapy are associated with higher mortality rates, longer-term mechanical ventilation, and excessive hospital costs. Traumatic injury places patients at particular risk for VAP, and efforts to perform accurate risk assessment and diagnostic confirmation should be focused in this population. Interpretation of meaningful information from the EMR at the bedside is complicated by high data volume, lack of integrated data displays and text-based clinical reports that may be reviewed only by manual search. This cumbersome data management strategy obscures the subtle signs of early infection.

The goal of our second project is to build NLP systems to identify patients who are developing critical illnesses in a manner timely enough for early treatment. As a first step, we have built a system that determines whether a patient has pneumonia based on the narrative text of the patient's chest X-ray reports.

To train and evaluate the system, we created a corpus of 1344 chest X-ray reports from our institution. Two annotators, one is a general surgeon and the other is a data analyst in a surgery department, read each report and determine whether the patient has pneumonia (PNA) and also what the clinical pulmonary infection score (CPIS) is for the patient. The CPIS is used to assist in the clinical diagnosis of VAP by predicting which patients will benefit from obtaining pulmonary cultures. The use of the CPIS is shown to result in fewer missed VAP episodes and can also prevent unnecessary antibiotic administration due to treatment of colonized patients.[2] There are three possible labels for CPIS: (1a) no infiltrate, (1b) diffuse infiltrate or atelectasis, and (1c) localized infiltrate. There are also three possible labels for PNA: (2a) no suspicion (negative class), (2b) suspicion of PNA, and (2c) probable PNA (positive class). The difference between the labels (2b) and (2c) is the certainty level on PNA. If there is enough evidence in a given report that indicates PNA, the report is labeled with (2c). If the evidence in the report is not enough to label it with (2c) but also not enough to rule out the possibility of PNA (2a), then it is labeled with (2b).

We used this corpus to train two classifiers, one for CPIS and the other for PNA). We did 5-fold cross validation. The accuracy of the CPIS classifier was 85.86%. The accuracy of the PNA classifier was 78.2% for the 3-way distinction, and the performance improved to 85.19% for the 2-way distinction when the two codes indicating suspicion of pneumonia, (2b) and (2c), were collapsed into a single class.

3.3 Pneumonia in the ICU Reports

With the introduction of comprehensive EMRs, all aspects of intensive care unit (ICU) care are now captured in both structured and free-text format. The existence of such data provides an opportunity to identify critical illness phenotypes and facilitate clinical and translational studies of large cohorts of critically ill patients, a task that would not be feasible using traditional screening/manual chart abstraction methods.

The goal of our third project is to build automated tools to identify critical illness phenotypes such as pneumonia (PNA) and model their progression based on the ICU reports. PNA can be classified further based on the context in which it occurs. Community acquired pneumonia (CAP) refers to pneumonia that occurs outside of the hospital setting; whereas hospital acquired pneumonia (HAP) refers to pneumonia which occurs after admission to the hospital. VAP is a special case of HAP, where the infection can be linked to the use of the ventilation machine.

Physician daily notes are a potentially rich source of clinical information indicating the presence of phenotypes such as pneumonia. In contrast to the narrow scope of information provided by radiology reports (e.g., chest X-ray reports), physician daily notes include text detailing patient narrative, physiologic, imaging, and laboratory data, and, finally, the physician's interpretation of these data. We hypothesized that by using physician notes such as admit notes, ICU progress notes, and discharge summaries, automated approaches that incorporate NLP and machine learning can accurately identify pneumonia in ICU settings.

To train and evaluate our PNA detection system, we created a corpus composed of ICU reports for 426 patients. An annotator with 6 years of experience as a research study nurse manually classified a patient as "positive" if the patient had pneumonia within the first 48 hours of ICU admission and as "negative" if the patient did not have pneumonia or the pneumonia was detected after the first 48 hours of ICU admission (66 cases positive for pneumonia and 360 cases negative for

[2] http://www.surgicalcriticalcare.net/Resources/CPIS.php

pneumonia). The annotation was per-patient. Because subjects in this dataset were admitted to the ICU from the emergency department as well as from other hospitals, cases of pneumonia included both CAP and HAP. Table 2 provides a summary of the characteristics of pneumonia.

CAUSES	
Bacteria: - *H. influenza* - *Strep pneumonia* - *Staph aureus* - Legionella species - Chlamydia species - *Pseudomonas aeruginosa*	Viruses: - Influenza - Parainfluenza Fungi: - Blastomycosis - Coccidiomycosis - Histoplasmosis
CLINICAL SIGNS AND SYMPTOMS	
Fever	Sputum production
Cough	Shortness of breath
Chest Pain	Malaise, fatigue
Abnormal white blood cell count	Muscle pains
RISK FACTORS	
Age > 65	
Immunosupression	
Recent antibiotic use	
Comorbid illnesses: HIV, Asthma, COPD, Renal Failure, CHF, Diabetes, Liver Disease, Cancer, Stroke	

Table 2. Characteristics of Pneumonia

Our dataset includes a total of 5313 reports from eight report types (admit note, ICU daily progress note, acute care daily progress note, interim summary, transfer/transition note, transfer summary, cardiology daily progress note, and discharge summary) for 426 patients. The total number of reports per patient ranged widely (median=8, interquartile range = 5-13, minimum =1, maximum=198). This is due to the high variability in the length of ICU stay. The distribution among the eight different report types is presented in Table 3. The first column of the table gives the number of reports for each report type and the second column gives the number of distinct patients who had the report type in the dataset.

REPORT TYPE	REPORT COUNT	PATIENT COUNT
ADMIT NOTES	481	280
ICU DAILY PROGRESS NOTE	2526	388
ACUTE CARE DAILY PROGRESS NOTE	1357	203
INTERIM SUMMARY	164	115
TRANSFER/TRANSITION NOTE	243	175
TRANSFER SUMMARY	18	18
CARDIOLOGY DAILY PROGRESS NOTE	133	17
DISCHARGE SUMMARY	391	350

Table 3. Statistics of the ICU corpus. Report Count: The number of reports with that report type; Patient Count: The number of distinct patients who had that report type.

In (Yetisgen-Yildiz, 2011b), we presented the preliminary results of the statistical system we built to identify PNA trained with this corpus. With 5-fold cross validation, our classifier achieved 58.3% precision, 42.4% recall, and 49.1% F1 for identifying patients with PNA. The classification accuracy was 86.4% and the specificity was 94.4%.[3]

4. Challenges

Given the nature of our annotation tasks, which relies on the medical expertise of annotators and requires protection of patients' privacy, the crowd-sourced annotation or community annotation schemata would not be applicable. Ideally, we would want to follow the traditional annotation schema, which has been proved to work well in numerous projects; however, we encounter several challenges due to some characteristics of our annotation tasks, and as a result, we have to make some changes to the traditional annotation schema.

4.1 Traditional Annotation Schema

In the traditional annotation schema, the annotation is done by a team consisting of the following members: project leader (l), guideline designers (d), linguistics / domain experts (e), annotators (a), and technical support (t). In addition, the team will ask its large research community (c) for suggestions, feedback and support.

Below is a common procedure for the traditional annotation schema, and the people who are in charge of each step are shown in parentheses:

1. Define annotation task based on the need of the community (l, c)
2. Select data to be annotated (l)
3. Write a detailed set of annotation guidelines (d, e)
4. Create good annotation tools (l, t)
5. Find and train annotators (l)
6. Annotate text
   a. Annotate text based on the guidelines (a)
   b. Revise annotation guidelines if needed (d, e)
   c. Monitor inter-annotator agreement and re-train annotators (l)
   d. Modify annotation based on the revised guidelines (a)
   e. Once some data have been annotated, train some NLP systems to pre-process the data to speed up annotation (l, t)
7. Release the corpus to the community (l)
8. Use the corpus to build various systems (c)
9. Find additional funding to extend the corpus, repeat some of the previous steps (l)

4.2 Characteristics of Clinical Annotation

Compared to most annotation projects in the general domain or the biomedical domain, our projects differ in several ways.

4.2.1 Annotation by Experts

For any annotation task on a non-general domain, having domain knowledge is helpful for the annotation team. The question is how much knowledge is required and how soon an annotator can acquire such knowledge. In our projects, medical expertise is a must for both design of the annotation guidelines and annotation itself, and it cannot be acquired quickly. As a result, we have to heavily rely on medical experts. We call this kind of

[3] Specificity is the negative (non-PNA) predictive value.

annotation “expert annotation”.

For instance, in the ICU corpus (C3), the annotator needs to go over all the ICU reports of a patient in order to determine whether the patient has pneumonia within 48 hours of admission to ICU. Very often, the ICU reports would not explicitly say whether or not the patient has pneumonia. The annotator, a research study nurse with six-year experience, has to use her medical expertise to determine whether the patient has any of the characteristics of the disease (see Table 2) and whether the identified characteristics are sufficient to make the call. For instance, when she sees the text “WBC: 15000 mcl” in a report, she knows that “WBC” stands for “white blood cell”, “mcl” stands for “microliter”, and the normal range of WBC count is 4,500-10,000 per microliter. So she knows that the text span indicates that the patient has “abnormal white blood cell count”, a symptom under “Clinical Signs and Symptoms”. Once she has found all the relevant cues in the text, she needs to then decide whether they are sufficient for her to label the patent as “having PNA”. All this domain knowledge cannot be acquired by a layman in a short period of time (say within a few months). Similarly, annotation guidelines such as the one in Table 4 for the chest X-ray corpus (C2) can be created and understood only by medical experts trained in a particular field.

4.2.2 Impact of Privacy Consideration

When annotating clinical data, privacy is an important concern. In addition to the requirement of getting IRB approval in advance, there are other ramifications; two examples are given here:

- The IRB review process can take a long time, and no one can work on the data before the IRB is approved. This leads to less flexibility in selecting the data set and choosing annotators. For instance, in the ICU project (C3), after the IRB approval, we got access to the records of the 426 patients listed on the IRB form, and we then realized that some patients missed important reports such as discharge summaries. But at that time, it was already too late to request records for additional patients, because that would require a new IRB approval, which could take additional time depending on the institution. Similarly, an annotator cannot work on a project unless the request of adding him/her to the project has been approved by the IRB.
- It is often very difficult for the annotation team to get approval to release the corpus to the research community. In the United States, HIPAA[4] provides guidelines for protecting patient information. HIPAA considers the data to be de-identified if the data is cleaned of seventeen categories of possible identifiers including personal health information (PHI) and any other information that may make it possible to identify the individual. Therefore, even if the corpus can be released, the de-identification process would make the corpus less useful for research purpose. If a corpus cannot be released, it becomes impossible for the community to benefit from the corpus and for the annotation team to get feedback from the community.

4.2.3 Impact of Legal Considerations

One characteristic of clinical domain is the concern about malpractice lawsuits. Let us use the radiology report corpus (C1) as an example. Poor communication has been found to be a causative factor in up to 80% of malpractice lawsuits involving radiologists (Levinson, 1994). In those lawsuits, the radiology report is often treated as an important medico-legal document. Given the legal aspect of the reports, it is common for a radiologist to use “hedging” in their reports (Wallis and McCoubrie, 2011), where “hedging” is “an evasive statement to avoid the risk of commitment” (Hall, 2000). Commonly used hedge phrases include *cannot exclude* and *not ruled out*.

From the perspective of annotation, hedging can be seen as ambiguity introduced by radiologists intentionally to keep certain information vague in order to protect themselves from potential lawsuits. If that information is related to what is being annotated, that could lead to annotation disagreement as annotators might interpret the radiologists’ intention differently. As an example, one annotator labeled the sentence “*If clinically indicated, pelvic ultrasound could be performed in 4 to 6 weeks to document resolution*” as critical recommendation, but the other annotator did not because he thought the author was hedging.

4.3 Effects on the Annotation Process

The differences discussed in the previous section affect the annotation process in several ways.

4.3.1 Roles of NLP Researchers

In a typical annotation project, NLP researchers often play a central role; they are team leaders, guideline designers, technical support staff, and users. They consult linguistic experts to write annotation guidelines; they hire and train annotators; they monitor inter-annotator agreement and re-train annotators; they build NLP systems to pre-process data to speed up annotation.

However, they play a more limited role in our clinical annotation projects because they lack the medical expertise to (1) design the task and write guidelines (e.g., what do the three labels for CPIS mean), (2) select relevant patient records, (3) select and train annotators, and (4) foresee potential legal ramifications. Those tasks often fall on the shoulders of physicians, who play the roles of domain experts, annotators, guideline designers, and sometimes users.

4.3.2 Guidelines

In all of our annotation projects, annotations are done by physicians. Physicians are not familiar with common practice of annotation, such as creating detailed annotation guidelines in advance and revising guidelines if necessary. They are accustomed to making decisions (e.g., reading ICU reports and determining whether a patient has pneumonia) based on their professional training. They might not believe that writing detailed guidelines is necessary, and even if they want to, turning

[4] Health Information Portability and Accountability Act (HIPAA), Section 164.514. Available at: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf

their medical knowledge into annotation guidelines can be very difficult. As a result, physicians often start annotation with no or very few annotation guidelines.

4.3.3 Finding and Training Annotators

Compared to annotators in a typical annotation task, physicians are much more expensive. They also have a very busy schedule and can spend very little time on annotation. Therefore, the common practice of training and re-training annotators, having several annotators work together to resolve disagreement, and having each annotator annotate 20 or more hours per week is all but impossible.

Training and re-training annotators is also difficult because the disagreement between physicians could be due to different interpretations related to their medical training. For instance, the sentence "*Correlation with ultrasound is advised*" is considered to be a critical recommendation by the radiologist but not by the clinician. (Re-)training them would mean that they have to change some long-held practice.

4.3.4 Process of Expert Annotation

Compared to traditional annotation schema, the team for clinical annotation is smaller, as the physicians (p) play the roles of guideline designers, annotators, and domain experts, whereas NLP researchers (n) provide technical support and build NLP systems using the corpus. The common process is as follows:

1. Define an annotation task based on the clinical needs (p)
2. Select data to be annotated (p)
3. Get IRB approval (p, n)
4. Write annotation guidelines (p)
5. Create good annotation tools (n)
6. Annotate text based on the guidelines and/or medical training (p)
7. Use the corpus to build various systems (n)
8. Test how well the systems meet the clinical needs (p)

One important lesson we have learned from our projects is that NLP researchers should get more involved in the process, especially in Steps 4 and 6, as demonstrated in the next section.

5. Strategies

While we have a lot of experience with annotation in the general domain and the biomedical domain, we had not worked on expert annotation previously. In this section, we summarize a few lessons we have learned from these projects.

5.1 Importance of Annotation Guidelines

It is well-known that having detailed annotation guidelines is crucial for training annotators and ensuring high inter-annotator agreement (IAA). But for expert annotation, the annotators, who are medical experts in our case, already know how to determine whether a patient has a certain illness as it is part of their routine job; furthermore, NLP researchers would not know how to train them since the annotators have more knowledge about the task. We therefore ask the question whether detailed annotation guidelines are still necessary at all, and how often physicians' judgments would agree with each other without the guidelines?

In order to answer the question, we obtained double annotation on all the 800 reports in corpus C1, and 100 of the 1344 reports in corpus C2.[5] For each corpus, we asked each annotator to do two rounds of annotation:

1. In the first round, there were no annotation guidelines other than the definition of critical recommendation for corpus C1, and the meaning of labels for corpus C2 (e.g., "2a" means "no suspicion of PNA"). Each annotator annotated the data independently from each other.
2. In the second round, the annotators went over the instances (an instance is a sentence in C1 and a report in C2) that received different labels in the first round and did the following:
   i. For C1, each annotator wrote a note to explain the rationale for his labeling; then he read the rationale written by the other annotator and relabeled the sentences if he agreed with the other annotator's rationale.
   ii. For C2, the two annotators discussed all the reports that received different labels and came up with a detailed set of guidelines (see Table 4 for the guidelines for CPIS). They then waited for a few days (so that they would be unlikely to remember the decisions on the 100 discussed reports) and re-annotated the reports based on the guidelines.

For the second round, we prefer (ii) over (i) as (ii) requires annotators to come up with detailed guidelines, which would be valuable when annotating new data, but we could not do that for corpus C1 due to the busy schedules of its two annotators.

1A: No Infiltrate
• The report includes information that neither diffuse nor localized infiltrate. The report could include edema or pleural effusion. • If there are extra pleural mentions in the report, they are not related to PNA.
1B: Diffuse Infiltrate or Atelectasis
• Atelectasis is more important than localized process that is consistent with infection. • Lobar collapse is consistent with atelectasis. • Multiple areas of opacity could fall under 1B. • If bi-basilar consolidation is present with bi-pleural effusion much more suggestive of atelectasis.
1C: Localized Infiltrate
• If one opacity is specifically highlighted and PNA or infection also mentioned in text, than this is more important than 1A and 1B.

Table 4: Annotation guidelines for determining CPIS labels in the chest X-ray corpus

With the two rounds of double annotations, we can calculate inter-annotator agreement (IAA) for each round. The results are shown in Tables 5-7. There are several observations. First, the IAA is pretty low for the first

[5] We did not do double annotation for Corpus C3 because we could not find another physician for the annotation task.

round, especially for the PNA labels in Table 7. Second, going through the second round with either (i) or (ii) improves the IAA significantly. Third, for the PNA labeling, the agreement is still low, 85%, even after the second round. All these indicate that solely relying on physicians' medical training is not sufficient in achieving a high IAA; creating detailed annotation guidelines and/or discussing examples with conflicting labels must be performed by physicians.

Round	A1	A2	Agreed	P/R/F	Kappa
1st	110	109	83	0.755/0.761/0.758	0.757
2nd	114	118	113	0.991/0.958/0.974	0.974

Table 5: IAA for the Radiology Corpus (C1). The corpus has 800 documents and 18,748 sentences in total. The "A1" and "A2" columns show the number of critical recommendation sentences (i.e., positive sentences) marked by the annotators; the "Agreed" column shows the number of positive sentences marked by both annotators; P/R/F scores are precision, recall, and F-score for identifying positive sentences when A2's annotation is treated as gold standard and A1's annotation is treated as system output; "kappa" is the kappa coefficient.

Round	A1	A2	Agreed	Acc	kappa
1st	13/59/28	15/74/11	12/52/6	70%	0.415
2nd	13/72/15	16/72/12	13/68/10	91%	0.797

Table 6: IAA on **CPIS** labeling for the 100 double annotated reports in the chest X-ray corpus (C2). x/y/z in each cell of the "A1", "A2", and "Agreed" columns are the numbers of reports with labels 1a, 1b, and 1c, respectively; "Acc" is the percentage of reports that receive the same CPIS label from the two annotators; "kappa" is the kappa coefficient.

Round	A1	A2	Agreed	Acc	kappa
1st	44/32/24	69/26/5	36/5/4	45%	0.085
2nd	67/19/15	67/32/1	66/18/1	85%	0.697

Table 7: IAA on **PNA** labels for the 100 double annotated reports in the chest X-ray corpus (C2). x/y/z in a cell of the "A1", "A2", and "Agreed" columns are the numbers of reports with labels 2a, 2b, and 2c, respectively; "Acc" is the percentage of reports that receive the same PNA label; "kappa" is the kappa coefficient.

5.2 Providing Additional Information

Another lesson we learned from this experience is that, in addition to the label of the instance, we should also ask annotators to mark additional information such as evidence or rationale. For instance, Corpus C3 currently includes only 426 yes/no labels, one for each patient. We do not know what kind of evidence the annotator has found in the reports to support her decision, and which reports the evidence comes from. Ideally, we would prefer to have the annotator mark the evidence in the report (e.g., the text "WBC: 15000 mcl" in the discharge summary) and link it to the characteristics of PNA listed in Table 2 (e.g., "Abnormal white blood cell count" under "Clinical Signs and Symptoms"). Marking such information will not only help NLP researchers to build better systems (e.g., the systems can learn what kinds of cues are relevant to the class label), but also help annotators to resolve any annotation disagreement.

When choosing granularity of annotation, one always need to consider the benefits of fine-grained annotation vs. the downside of increased annotation time. For corpus C3, in order to give correct PNA labels, the annotators have to read the whole reports and look for those cues; as such, highlighting relevant text spans and clicking some buttons to link cues to some pre-defined characteristics would not substantially increase annotation time. The additional time is well spent since a patient has tens to hundreds of ICU reports, and therefore knowing where the cues come from will greatly reduce the number of features that an NLP classifier has to consider. We plan to include such additional information in the next stage of the project. For corpus C2, we also plan to mark the text span, although the benefits are less than in C3, because the reports in C2 are much shorter and the annotation is already at the report level, not the patient level.

5.3 Time Commitment from Physicians

All the projects discussed in Section 3 were initiated by our physicians. They are very interested in building NLP systems to meet their clinical needs. However, because they are not familiar with annotation process, they often underestimate the amount of time required for annotation, guideline designs, and other related activities. Their busy schedule at the hospital often limits the amount of time they can spend on the project.

To address this problem, we, the NLP researchers, should explain to the physicians what the annotation process looks like and why having detailed annotation guidelines and monitoring IAA are important. We should also provide them a good estimate of time commitment that will be required to complete the project. They can then make an informative decision on whether they are able to devote enough time to the project.

5.4 Early Involvement of NLP Researchers

Although it may be true that NLP researchers play a minor role in expert annotation, they should still get involved in the annotation process as early as possible. Despite their lack of medical training, they can help physicians in each step of the annotation process described in Section 4.3.4. For instance, they can calculate IAA and convince physicians to write detailed guidelines; they can inform physicians what kind of additional information would be beneficial to add; they can help physicians to decide how big the corpus needs to be; they can pre-process the data to filter out noisy data.

6. Conclusion

In this paper, we discuss three corpora that we created for clinical NLP projects. Unlike most of the previous annotation projects, these corpora require expert annotation. Our studies show that, without detailed guidelines and/or discussion, the annotation agreement among experts is low, indicating medical training itself is not sufficient for high-quality annotation. Although NLP researchers lack medical training and therefore play a minor role in guideline designs and annotation, their early involvement is important for the success of annotation.

7. References

Chapman W.W., Cohen K.B. (2009). Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics. 42(5):757-759.

Chapman W.W., Nadkarni P.M., Hirschman L., D'Avolio L.W., Savova G.K., Uzuner O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. JAMIA[6]. 18:540-543.

Fiszman M., Chapman W.W., Aronsky D., Evans R.S., Haug P.J. (2000). Automatic detection of acute bacterial pneumonia from chest x-ray reports. JAMIA, 7(6):593-604.

Franzen K., Eriksson G., Olsson F., Lidin L.A.P., Coster J. (2002). Protein names and how to find them. International Journal of Medical Informatics. 67(1-3):49-61.

Hajic, J. (1999). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajicova (Ed.), Issues of valency and meaning. Studies in honour of Jarmila Panevova. Prague, Czech Republic: Charles University Press.

Hall F. (2000). Language of the radiology report: primer for residents and wayward radiologists. AJR American Journal of Roentgenol. 175:1239-1242.

Kim J.D., Ohta T., Tateisi Y., Tsujii J. (2003). GENIA corpus–semantically annotated corpus for bio-text mining. Bioinformatics. 19:Suppl 1:180-2.

Kulick S., Bies A., Liberman M., Mandel M., McDonald R., Palmer M, Schein, Ungar L. (2004). Integrated annotation for biomedical information extraction. In Proceedings of HLT-NAACL workshop BioLink 2004, Linking Biomedical Literature, Ontologies, and Databases. pp. 61-8.

Levinson W. (1994). Physician-patient communication: a key to malpractice prevention. Journal of the American Medical Association (JAMA). 272:1619-1620.

Lucey L.L., Kushner D.C. (2010). The ACR Guideline on Communication: To Be or Not to Be, That Is the Question. Journal of the American College of Radiology. 7(2): 109-114.

Maamouri M and Bies A. (2004). Developing an Arabic treebank: methods, guidelines, procedures, and tools. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.

Marcus M., Marcinkiewicz M.A., and Santorini B. (1993). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, 19(2).

Miltsakaki E., Prasad R., Joshi A., and Webber B. (2004). The Penn Discourse TreeBank. In Proc. of LREC.

Pakhomov S.V., Coden A., Chute C.G. (2006). Developing a corpus of clinical notes manually annotated for part of speech. International Journal of Medical Informatics. 75(6):418-429.

Palmer M., Gildea D., and Kingsbury P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics. 31(1): 71-106.

Roberts A., Gaizauskas R., Hepple M., Davis N., Demetriou G., Guo Y., Kola J., Roberts I., Setzer A., Tapuria A., Wheeldin B. (2007). The CLEF Corpus: Semantic Annotation of Clinical Text. In Proceedings of AMIA Annual Symposium. pp. 625-629.

Roberts A., Gaizauskas R., Hepple M., Demetriou G., Guo Y., Roberts I., and Setzer A. (2009) Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics. 42: 950-966.

Savova G.K., Chapman W.W., Zheng J., Crowley R.S. (2011). Anaphoric relations in the clinical narrative: corpus creation. JAMIA. 18:459-465.

Snow R., O'Connor B., Jurafsky D., Ng A.Y. (2008). Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of EMNLP'08. pp.254-263.

South B.R., Shen S., Jones M., Garvin J., Samore M.H., Chapman W.W., and Gundlapalli A.V. (2009). Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics. 10(Suppl 9):S12.

South B.R., Shen S., Barrus R., DuVall S.L., Uzuner O., Weir C. (2011). Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In Proceedings of the AMIA Annual Symposium. pp. 1243-1251.

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

Towbin A.J., Hall S., Moskovitz J., Johnson N.D., Donnelly L.F. (2011) Creating a comprehensive customer service program to help convey critical and acute results of radiology studies. AJR American Journal of Roentgenology. 196(1):W48-51.

Uzuner O., Goldstein I., Luo Y., Kohane I. (2008) Identifying patient smoking status from medical discharge records. JAMIA. 15(1):14-24.

Uzuner O. (2009). Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association. 16(4):561-570.

Uzuner O, Solti I, Xia F., and Cadag E. (2010). Community Annotation Experiment for Ground Truth Generation for the i2b2 Medication Challenge. JAMIA. 17:519-523.

Uzuner O, South B.R., Shen S., DuVall S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. JAMIA. 18(5): 552-556.

Yetisgen-Yildiz M, Solti I., Xia F., Halgrim S.R. (2010) Preliminary Experiments with Amazon's Mechanical Turk for Annotating Medical Named Entities. In Proc. of Creating Speech and Language Data with Amazon's Mechanical Turk Workshop of NAACL'2010.

Yetisgen-Yildiz M, Gunn ML, Xia F, Payne T. (2011a). Automatic identification of critical follow-up recommendation sentences in radiology reports. In Proceedings of AMIA Annual Symposium. pp. 1593-1602.

Yetisgen-Yildiz M, Glavan BJ, Xia F, Vanderwende L, Wurfel MM. (2011b). Identifying Patients with Pneumonia from Free-Text Intensive Care Unit Reports. In Proceedings of Learning from Unstructured Clinical Text Workshop of ICML'2011.

Wallis A., McCoubrie P. (2011). The radiology report – Are we getting the message across? Clinical Radiology. 66(11):1015-1022.

Xia F., Palmer M., Xue N., Okurowski M.E., Kovarik J., Huang S., Kroch T., and Marcus M. (2000). Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In Proceedings of LREC.

[6] *JAMIA* stands for Journal of American Medical Informatics Association.

The Journal of the Swedish Medical Association - a Corpus Resource for Biomedical Text Mining in Swedish

Dimitrios Kokkinakis
Center of Language Technology and Språkbanken
University of Gothenburg, Sweden
Box 200, 405 30 Gothenburg
E-mail: dimitrios.kokkinakis@svenska.gu.se

Abstract

Biomedical text mining applications are largely dependent on high quality knowledge resources. Traditionally, these resources include lexical databases, terminologies, nomenclatures and ontologies and, during the last decade, also corpora of various sizes, variety and diversity. Some of these corpora are annotated with an expanding range of information types and metadata while others become available with a minimal set of annotations. It is also of great importance that biomedical corpora for lesser-spoken languages also get developed. This is required in order to support and facilitate implementation of practical applications for such languages and to stimulate the development of language technology research and innovation infrastructures in the domain. This paper provides a description of a Swedish biomedical corpus based on the electronic editions of the *Journal of the Swedish Medical Association* "Läkartidningen" of the years 1996-2010. The corpus consists of a variety of documents that can be related to different medical domains, developed as a response to the increasing needs for large and reliable medical information for Swedish biomedical Natural Language Processing (NLP). The corpus has been structurally annotated with a minimal set of meta information and automatically indexed with the *Swedish Systematized Nomenclature of Medicine -- Clinical Terms* (SNOMED CT).

Keywords: Swedish biomedical text corpus, Structural annotation, SNOMED CT, Semantic annotation, META-NET

1. Introduction

With the information overload in the life sciences there is an increasing need for corpora, raw or preferable annotated, which is the driving force for data-driven language processing applications and the empirical approach to language study in various domains. At the same time, there is an overwhelming and growing amount of data and information on the web, easily accessible but not as easily controllable as with respect to accuracy, trustworthiness and openness. Nonetheless, exceptions do exist, and the most prominent example is the PubMed/MEDLINE, an exponentially growing database of abstracts, for, primarily, English, that has been the *de facto* standard for acquiring input documents for a large number of biomedical NLP-related projects and initiatives. Usually, such projects are based on relatively small subsets from PubMed/MEDLINE and in narrow subdomains since manual annotation and curation are time-consuming and costly. For instance, a number of corpora, such as the *GENIA corpus* (Kim *et al.*, 2003), extensively used in the biomedical field, is heavily based on the PubMed's content.

2. Background

The number of biomedical corpora increase steadily. Most of the corpora are annotated with an expanding range of simple and complex information types, such as named entities, and relations that hold between them, coreference and various event and other higher level discourse structures. Here, we provide a brief description of some of these corpora and from a biomedical NLP point of view, which implies that we consciously ignore description of e.g. clinical data. Undoubtedly, the most widely used is the GENIA corpus, a fully annotated material of 2,000 PubMed abstracts, with semantically-oriented markup, such as named entities. GENETAG (Tanabe *et al.*, 2005) is a corpus of 20K PubMed sentences for gene/protein entity recognition; 15K of these sentences were used for the BioCreAtIvE[1] Task 1A Competition. The AIMed corpus (Bunescu *et al.*, 2005) is a corpus of 200 PubMed abstracts, created for protein-protein interaction extraction method comparison. The abstracts were manually annotated for interactions between human genes and proteins. In addition to the 200 abstracts, further 30, without protein-protein interactions, were added to the corpus as negative examples. The PennBioIE CYP corpus contains 1,100 PubMed abstracts (non-exhaustively) annotated for 5 types of named entity on the inhibition of cytochrome P450 enzymes and the PennBioIE oncology which consists of 1,414 PubMed abstracts on cancer, concentrating on molecular genetics. (Mandel, 2006), BioInfer (Bio Information Extraction Resource), is an annotated corpus that contains 1,100 sentences from abstracts of biomedical research articles annotated for relationships, named entities, as well as syntactic dependencies; *cf.* Pyysalo *et al.*, 2007. The GREC corpus (Thompson *et al.*, 2009) is yet another semantically annotated corpus of 240 PubMed abstracts (167 on the subject of E. coli species and 73 on the subject of the Human species) which is intended for training IE systems and/or resources which are used to extract events. In a larger scale, Rebholz-Schuhmann *et al.* (2010) describe an effort to annotate 150,000 PubMed abstracts on immunology, with various semantic entity types in the course of two annotation challenges in the framework of the Collaborative Annotation of a Large Biomedical

[1] http://biocreative.sourceforge.net/

Corpus project (CALBC). Finally, the BioMed Central's open access full-text corpus for text mining research contains a growing amount of articles (as of the February of 2012 BioMed Central has published 117,925 peer-reviewed articles) all of which are covered by our open access license agreement which allows free distribution and re-use of the full-text article, including the highly structured XML version.

3. The Journal of the Swedish Medical Association: *Läkartidningen*

The Swedish Medical Association's Journal ("*Svenska Läkartidningen*", LT), has been for over a century the main source of reliable medical knowledge of the 1903 established General Swedish Medical Association, now simply referred to as Medical Association, "*Läkarförbundet*" *(cf.* Eklöf, 2000). Over the years LT has emerged as an authentic, reliable national knowledge resource. It is now widely used as a source of knowledge for up-to-date scientific medical information not only by the health care system's different staff groups and academic researchers but also by the general public who want a reliable basis for their own reflections on health related topics or wish to acquaint themselves with the new findings and developments of the medical knowledge domain in their native language, Swedish. The LT archive is the largest Swedish-language source material in medicine. LT is also an important point of reference, a genuine language and source of inspiration for terminologists, linguists and specialized language professionals who want e.g., to determine how medical terms and concepts are used in authentic medical texts. The LT's material can fulfill a variety of scientific, societal and technological needs. Authentic textual data are for instance fundamental for empirical studies in terminography, language technology, and linguistics and there is a growing need for such high quality data that can strengthen the national resource infrastructure.

The breadth of the material provides a suitable platform for studies in an array of disciplines that can fulfill various research interests; for instance about diagnoses, treatment protocols and outcomes, in a broader perspective over a long period of health care. Electronic editions of the journal are made available since 1996 and it is that period up to the end of 2010 that this paper is describing. The electronically accessible part of the archive in print quality (as of 1996, vol. 93) is accessible in various ways. One offered option allows for queries based on the use of keywords taken from a comprehensive concept hierarchy, Medical Subject Headings (MeSH), used for indexing journal articles and books in the life sciences. MeSH concepts have been manually assigned in advance to each electronic article <http://ltarkiv.lakartidningen.se/>. Today, the LT's digital archive (1996-) consists of different types of text articles and short news of scientific nature. There are currently over 30,000 such articles electronically available with valuable scientific and clinical information in various disciplines, health economic evaluations and analyses, medical historical views, pharmaceutical studies and medical language issues and new scientific findings etc. The current electronic archive spans all genres and medical disciplines, one of the reasons which makes it unique and usable for both a broad audience and specialists. For instance, the electronic editions of LT have been already used for quality assessment of the Swedish translation of the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT); *cf.* Kokkinakis & Gerdin, 2010.

4. Current Status of the Journal's Content

This section provides a detailed description of the Journals content, which covers vol. 93-107 (1996-2010).

4.1 Corpus Processing and Harmonization

Volumes 93-102 (1996-2005) of LT were only available as pdf files, while since 2005 the content is also published in formats such as *.xml* and *.html*, which are easier to process. Although the non-pdf editions of the Journal are rather fairly unproblematic for NLP processing, the pdf files pose certain difficulties due to the complexity of the layout of the journal’s pages and the different pdf versions that the material is encoded in. Therefore we decided to harmonize all data. All material has been transformed to a unified UTF-8 text-format.

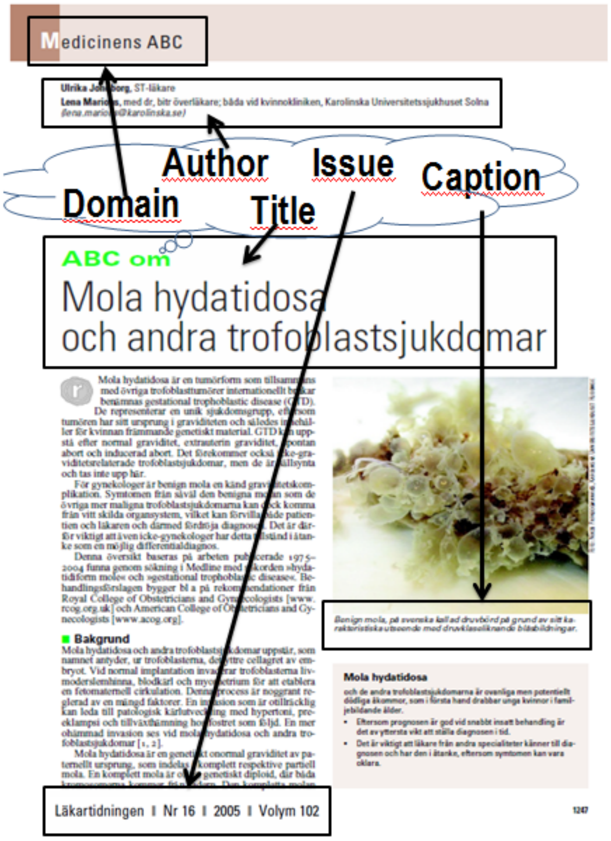


Figure 1. A rather typical pdf page from 2005;102:16.

The extraction of the text from the PDF files was made in an automatic fashion, using the ABBYY PDF Transformer 2.0, with manual verification. Our aim was to preserve as much as possible of the logical text flow and eliminate the risk for losing valuable information such as basic metadata, e.g. each article’s title and publication details. Figure 1 shows a typical page from a paper published issue in which relevant structural information is explicitly marked (and subsequently extracted) namely *domain*, *author*, *title/header*, *issue, publication date* as well as *table* and *figure captions*.

4.2 Corpus Description

The basic preprocessing steps are tokenization and sentence identification. The whole material is tokenized and segmented into sentences, using adapted generic NLP tools. Since some of the texts were of very technical, certain modifications were made to the tokenizer in order to properly handle erroneously tokenised special cases such as: *...ämnet NKK (4-(methylnitrosamino)-1-(3-pyri-dyl)-1-butanone) i urinen...* (i.e. "...the NKK substance (4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone) in urine...").

A part of the processing is also the automatic annotation of the corpus with the Swedish SNOMED CT, the Systematized Nomenclature of Medicine Clinical Terms. SNOMED CT provides a common language that enables consistency in capturing, storing, retrieving, sharing and aggregating health data across specialties and sites of care. SNOMED CT provides codes and concept definitions for most clinical areas. According to the international release of Jan. 2012, it includes more than 315,000 active concepts, where each concept is claimed to have formal ontological definitions. SNOMED CT concepts are organized into 18 top-level hierarchies, such as *Body structure* and *Clinical Finding*, each subdivided into several sub-hierarchies. More detailed information about SNOMED CT can be found at the International Health Terminology Standards Development Organisation's web site, IHTSDO, at: <http://www. ihtsdo.org/snomed-ct/>).

The annotation using SNOMED CT facilitates the rapid development of high quality, relevant subcorpora of a particular domain or topic using the assigned concepts which can be used as indexes of the articles. Figure 2, below, shows some characteristics of the corpus which is currently comprised of 30,002 different articles and 28,2 million tokens. Since 2006 there is also a possibility to comment the electronically published articles. This is a rapidly increasing trend that can be observed in the material, from 13 commented articles in 2006 to 405 in 2010. Available comments are suitably annotated and saved under the article they refer to. Since the material is tokenized it is also rather trivial to generate different types of statistics based on its content such as the longest words without a hyphen (e.g. *videoradioultrasonomagnetografonuklearmedicin*; 45 characters; vol. 99:(17): 1959); the longest words with a hyphen (e.g. *hallucination-cenestopati-depersonalisationssyndromet*; 53 characters, vol. 104:(30-31): 2152); or the top-5 most frequent common nouns: *procent* ('percent'; 32641), *patienter* ('patients'; 31904), *läkare* ('doctor/physician'; 21137), *behandling* ('treatment'; 18497) and *patienten* ('the patient'; 17920). Table 1 shows some more descriptive details of the nature of the content (1996-2010) that the current material covers. In table 1, *Real Words* is the total number of tokens, except punctuation and numerical data (alone or in combination) as well as emails and URLs; *Real Unique* is the same as the previous but here all tokens are normalized with respect to case, repeated/duplicate tokens are counted as one token.

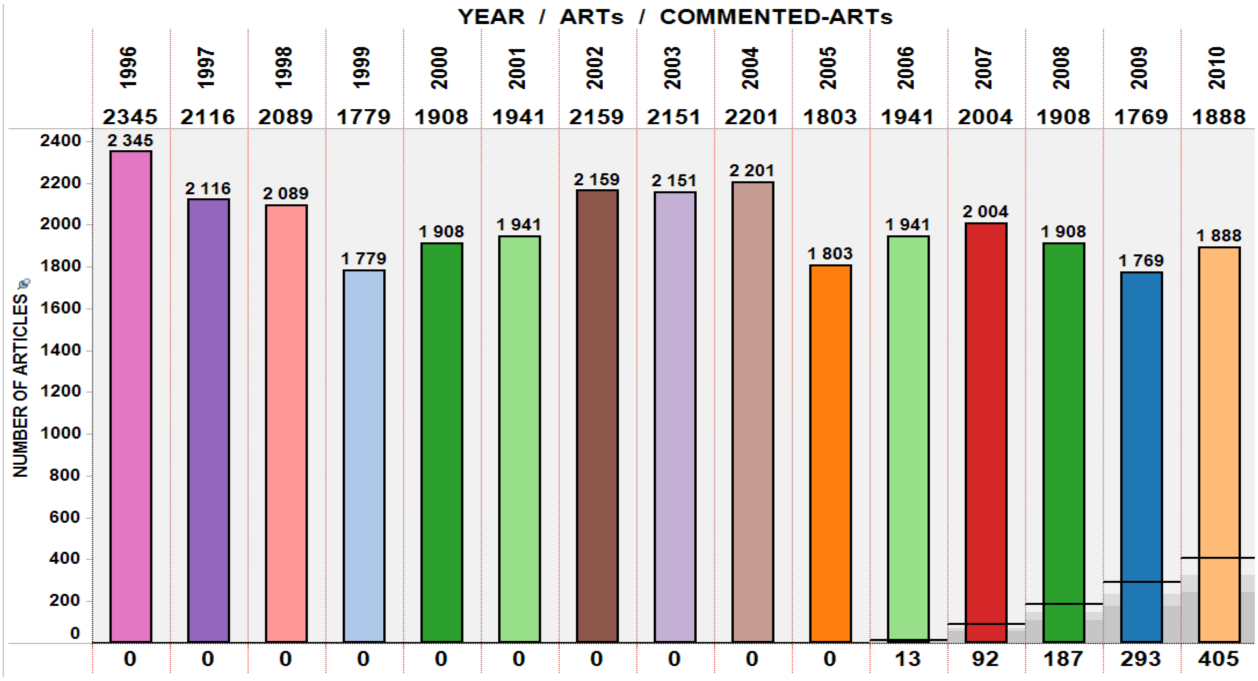


Figure 2. Some characteristics of the Swedish Medical Association Journal corpus, the distribution of articles per year and also the number of commented articles (x-axis).

ANL, AWL and *ASL* are the average noun, all words' and sentence length respectively while ELW is the proportion of exceptionally long words =>14 chars (note that Swedish is a compounding language so the proportion of long words *is* expected to be relatively high, for comparison reasons, the proportion of ELW in daily newspaper corpora is less than 2.8%).

Year	"Real" words	"Real" unique	ANL	AWL	ASL	ELW
1996	1772304	113002	9.34	5.91	18.8	5.2%
1997	1741476	111230	9.33	5.86	19.1	5.1%
1998	1934678	120135	9.18	5.77	19.4	4.7%
1999	1826502	115257	9.23	5.81	19.6	4.8%
2000	1761365	109921	9.36	5.87	19.9	5.1%
2001	1842160	111847	9.22	5.83	19.7	4.8%
2003	1772555	104522	9.29	5.86	19.6	5%
2003	1543350	94760	9.38	5.87	19.5	5.1%
2004	1603939	102560	9.36	5.92	19.3	5.2%
2005	1320779	89357	9.32	5.91	19.4	5.2%
2006	1388193	90846	9.34	5.92	19.7	5.2%
2007	1431400	93535	9.36	5.96	19.5	5.3%
2008	1520973	99292	9.32	5.94	19.9	5.3%
2009	1478623	96044	9.31	5.93	20.0	5.2%
2010	1468252	98974	9.29	5.98	20.1	5.3%
ALL	**24406549**	**551456**	**9.3**	**5.88**	**19.56**	**5.1%**

Table 1. Descriptive characteristics of the corpus.

Also, the annotation with SNOMED CT provides a good opportunity to accurately measure the presence of various terms since effort has been put to generate variant forms and near synonyms which can be easily linked and aggregated to their concept id. Figure 3 below shows the distribution of the terms *diabetes mellitus type 1* and *type 2* (id 44054006 and id 46635009, second and third line from the top) as well as the *gestational diabetes*, *diabetes insipidus* and *diabetes* id 73211009 (general mentions, which is actually the top line). In this aggregated view, the line of e.g. *diabetes mellitus type 2* also incorporates variant mentions such as: *diabetes type 2*; *type II-diabetes* and *type 2 DM.*

4.3 Subcorpora Extraction

Based on the SNOMED CT annotated version of the data we can now easily create "focused" subcorpora (extract article sets) that fulfil certain criteria, since each term mention is automatically assigned a number of attributes in a consistent manner; *cf.* Kokkinakis, (2011). For instance, a synonym to *leprosy* in Swedish is *spetälska*, thus a text occurrence of this term is annotated as: <snomed c="disorder" h="81004002" o="lepra" f="new">spetälska</snomed>, where *c* is a hierarchy, *h* the concept id, *o* the recommended term and *f* the result of the recognition process; here a *new* implies that the terms is taken from a synonym term list, other values could be e.g. *inflection*, for an inflected variant of a recommended term or *acronym*, if the term is an "unofficial" short variant of a recommended SNOMED term. For the subcorpora extraction, several methods can be used, e.g. clustering based on the annotations; by calculating the co-occurrence or frequency of concept ids based on tf*idf normalization or other relevant scores and measures.

5. Conclusions

This paper provided a description of the electronic editions of the Journal of the Swedish Medical Association, *Läkartidningen*. Although the journal has been the main source of reliable medical knowledge of the 1903 established Swedish Medical Association for over a century, only paper editions exist for the issues printed during 1903-1995. It would have been a great source of research if all issues could one day be available electronically. In particular, since the archive is the largest Swedish-language source material in medicine and also an important point of reference, a genuine language and source of inspiration for terminologists and specialized language professionals who want e.g. to determine how medical terms and concepts are used in authentic medical texts in a diachronic perspective. Currently, the journal is widely used as a source of knowledge for up-to-date scientific medical information not only by the health care system's different staff groups and academic researchers

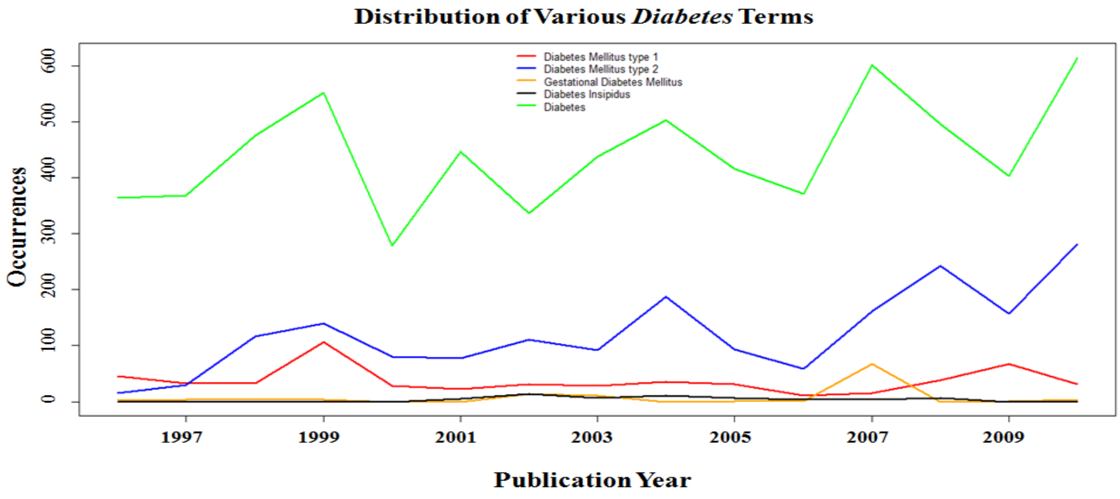


Figure 3. Distribution of diabetes-related terms in the corpus.

but even by the general public who want a reliable basis for their own reflections on health related topics or wish to acquaint themselves with the latest developments of the medical domain in their native language, Swedish.

In order to promote the interoperability and (re)use of the described resource for biomedical NLP-related research we have also started to describe its content according to the META-NET schema. META-NET is dedicated to building the technological foundations of a multilingual European information society and to its aim is to push forward research to allow a rapid expansion of language technologies <http://www.meta-net.eu/>.

6. Acknowledgements

The author wishes to thank the editors of the Journal of the Swedish Medical Association for the availability of the electronic versions of the magazine.

7. References

Razvan Bunescu, Ruifang Gea, Rohit J. Katea, Edward M. Marcotteb, Raymond J. Mooneya, Arun K. Ramanib and Yuk Wah Wong. (2005). Comparative Experiments on Learning Information Extractors for Proteins and their Interactions. *J AI Med*. 33:139–155.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi and Jun'ichi Tsujii. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining**.** *Bioinformatics*. 19(S1). Pp. i180-i182, OUP.

Motzi Eklöf. (2000). Läkarens Ethos - Studier i den svenska läkarkårens identiteter, intressen och ideal 1890-1960. *Linköping Studies in Arts & Science*, Nr 216. <http://liu.diva-portal.org/smash/get/diva2:22817/FULLTEXT01> (in Swedish)

Dimitrios Kokkinakis and Ulla Gerdin. (2010). A Swedish Scientific Medical Corpus for Terminology Management and Linguistic Exploration. 7th international Conf. on Language Resources and Evaluation (LREC). Pp. 2330-2335. Malta.

Dimitrios Kokkinakis (2011). What is the Coverage of SNOMED CT® on Scientific Medical Corpora? XXIII MIE. European Federation for Medical Informatics. Vol. 169: 814-818. Oslo, Norway.

Mark A. Mandel. (2006). Integrated Annotation of Biomedical Text: Creating the PennBioIE Corpus. Text Mining, Ontologies and NLP in Biomedicine. Manchester, UK.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen and Tapio Salakoski. (2007). BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8:50.

Dietrich Rebholz-Schuhmann, Antonio José Yepes, Erok M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger and Udo Hahn. (2010). CALBC Silver Standard Corpus. *J of Bioinf and Comp. Biology.* 8:1. Pp 163-179.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten and W John Wilbur. (2005). GENETAG: a Tagged Corpus for Gene/Protein NER. *BMC Bioinformatics*. 6(S 1):S3.

Paul Thompson, Iqbal SA., McNaught J. and Ananiadou S. (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics* 10:349

Releasing a Swedish Clinical Corpus after Removing all Words – De-identification Experiments with Conditional Random Fields and Random Forests

Hercules Dalianis and Henrik Boström
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, 164 40 Kista
E-mail: hercules@dsv.su.se, henrik.bostrom@dsv.su.se

Abstract

Patient records contain valuable information in the form of both structured data and free text; however this information is sensitive since it can reveal the identity of patients. In order to allow new methods and techniques to be developed and evaluated on real world clinical data without revealing such sensitive information, researchers could be given access to de-identified records without protected health information (PHI), such as names, telephone numbers, and so on. One approach to minimizing the risk of revealing PHI when releasing text corpora from such records is to include only features of the words instead of the words themselves. Such features may include parts of speech, word length, and so on from which the sensitive information cannot be derived. In order to investigate what performance losses can be expected when replacing specific words with features, an experiment with two state-of-the-art machine learning methods, conditional random fields and random forests, is presented, comparing their ability to support de-identification, using the Stockholm EPR PHI corpus as a benchmark test. The results indicate severe performance losses when the actual words are removed, leading to the conclusion that the chosen features are not sufficient for the suggested approach to be viable.

Keywords: de-identification, conditional random fields, random forests, Swedish clinical text

1. Introduction

A huge amount of clinical texts are produced today in electronic patient record systems where clinical personnel enter the status of the patient, including symptoms, medication, blood values, x-ray pictures, diagnosis codes, and so on. In addition to supporting the care of the individual patients, this information can potentially have a high value for research. However, for reasons of confidentiality, this type of information cannot easily be made accessible to researchers outside the clinics.

The electronic documents contain personal information about the patient, including details of relatives, phone numbers, addresses, and so on. This type of information, which can potentially reveal the identity of a patient, is often referred to as Protected Health Information (PHI). Obviously, it would be a great advantage if the information in the electronic patient records could be made accessible for research and development purposes without revealing the identity of the patients and their relatives. To effectively and efficiently de-identify patient records, both human and computer resources are required. However, as stated by Ohm (2009), even if a clinical text is fully de-identified, often it can still be easily be re-identified. The main question is whether or not one can achieve 100 percent de-identification while still keeping useful information for research and development purposes. One such approach would be to remove *all* words, keeping only features of the words from which the sensitive information cannot be derived.

2. Previous Research

A good overview of the area of de-identification of clinical documents can be found in Meystre et al. (2010), including a discussion of the limitations of the de-identification systems as well as conclusions about which methods and approaches are most advantageous for de-identification of clinical documents. The best systems developed for clinical text written in English achieve average precision, recall, and F-scores of between 0.90 and 0.96 with the standard 18 PHI-classes (HIPAA, 2003). However, Meystre et al. (2010) do not mention the amount of over-scrubbing (that is, removing too much information) of clinical findings and symptoms as well as common words. The available clinical corpora that can be used for research are all de-identified by computers in conjunction with manual scrubbing and for that reason are not particularly large, that is, rarely larger than 400 000 tokens. To gain access to such data, users have to sign confidentiality agreements. For details about the different available clinical corpora, see Alfalahi (2011) and Alfalahi et al. (2012).

Velupillai et al. (2009) describe a set of patient records written in Swedish that has been annotated by three different annotators for de-identification purposes. These patient records encompass 100 patient records (with a distribution of 50 percent men and 50 percent women) from five different clinics: pain, orthopaedic, oral, and maxillofacial surgery, and diet, containing 380 000 tokens. Later, a consensus of the three sets of annotations was created (Dalianis & Velupillai 2010). This set is referred to as the Stockholm EPR PHI corpus and it contains 4 480 (consensus) annotation instances distributed over the eight annotation (PHI) classes; *Age, Date_Part*, *Full_Date*, *First_Name*, *Last_Name*, *Health_Care_Unit, Location*, and *Phone_Number*. These correspond to 1.6 percent of the total set of tokens. Using the Stanford CRF (Conditional Random Fields) NER algorithm (Finkel et al. 2005), an F-score of 0.80 with a precision of 0.90 and recall of 0.72 was obtained (Velupillai & Dalianis 2010). Kokkinakis and Thurin (2007) obtained 0.97 precision and 0.89 recall when de-identifying 200 discharge letters written in Swedish using rule-based methods and name lists.

Better results are required, particularly with respect to higher recall, since for privacy reasons it is important not to miss any sensitive information.

3. Method and Materials

We will compare two state-of-the-art machine learning methods, conditional random fields (CRF; Lafferty et al. 2001) and random forests (Breiman 2001), regarding their ability to support de-identification. CRF is a machine learning method for segmenting and labelling sequence data. In this study, we employ the CRF++ implementation (CRF++ 2011), which in addition to using the words themselves as features may also consider other features, including part-of-speech (POS) tags, word length, and other structural and morphological information.

The random forest algorithm (Breiman 2001) generates a set of classification trees (Breiman et al. 1984), while incorporating randomness both in the selection of training examples and in the selection of features to consider when generating each individual tree. The former is done by employing bootstrap aggregating, or bagging (Breiman 1996), which works by randomly selecting n examples with replacements from the initial set of n training examples. Furthermore, when generating each tree in the forest, only a small randomly selected subset of all available input features is considered at each node in the tree. Random forests are widely considered to be among the most competitive and robust of current methods of predictive data mining (Caruana & Niculescu-Mizil 2006). The implementation that is used in the study is a parallel version that has been developed in Erlang (Boström 2011). The random forest algorithm is provided with the same features as CRF++, except that the words in the clinical texts have been excluded.

These methods have been applied on a clinical text called the Stockholm EPR PHI corpus[1] (Dalianis & Velupillai 2010). The corpus can be considered as a stream of tokens, some of which are of course regular words and sentences. Following standard approaches (see, e.g., Olsson 2008), we have chosen to represent words using the following 14 features.:

i) Is the token alpha numeric?
ii) Is it numerical?
iii) Does it have an initial capital letter?
iv) What is the POS tag two tokens before the token?
v) What is the POS tag one token before the token?
vi) What is the POS tag of the specific token?
vii) What is the POS tag one token after the token?
viii) What is the POS tag two tokens after the token?
ix) What is the token length two tokens before the token?
x) What is the token length one token before?
xi) What is the specific token length?
xii) What is the token length one token after the token?
xiii) What is the token length two tokens after?
xiv) What is the PHI class of the token?

The last (no. xiv) of the 14 features hence contains the target (output) value, which is typically unknown in novel (untagged) documents. As mentioned above, there are eight possible annotation classes, which, together with the non-PHI value, result in nine possible class values for the target feature.

For the CRF++, we used the word itself as a feature, which is standard for CRF, but also included the same feature set as for the random forest algorithm. CRF++ has a built-in function to use a window of up to four tokens before and up to four tokens after the token that is to be classified. This built-in window function therefore makes it possible to derive the 14 features above from the following limited set:

i) Is the token alpha numeric?
ii) Is it numerical?
iii) Does it have an initial capital letter?
iv) What is the POS tag of the specific token?
v) What is the specific token length?
vi) What is the PHI class of the token?

As a comparison we also used CRF++ without words but with the POS tags as features.

We also selected the maximum window size, that is, four tokens before and four tokens after the token to be classified, giving a total window size of nine. This turned out to give the best results in preliminary experiments. Tenfold cross validation is used in the evaluation (Kohavi 1995).

The differences between the approach used here, applying CRF++ with POS tags as well as 14 features, and the approach in Dalianis and Velupillai (2010) using Stanford CRF NER are that we only used the words and the PHI as features and that random forest was not used in Dalianis and Velupillai (2010).

4. Results

In Table 1, it can be observed that when removing the actual words, the performance of CRF++ drops radically in most cases with respect to all three criteria; precision, recall and F-score. Although random forests without words in several cases is able to obtain a higher precision than CRF++ with words, this carries over to the F-score

[1] The study was carried out after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

Tenfold cross evaluation		**CRF++ (with words and 14 features)**			**CRF++ (w/o words and only POS tags)**			**Random forest (w/o words and 14 features)**		
Classes	**Instances**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Age	56	0.860	0.704	0.774	**0.917**	0.650	0.761	**0.928**	**0.741**	0.824
Date_Part	711	**0.872**	**0.872**	0.872	0.724	0.745	0.735	0.839	0.663	0.741
Full_Date	551	0.841	**0.880**	0.860	0.570	0.451	0.503	**0.917**	0.819	0.865
First_Name	923	**0.918**	**0.763**	0.834	**0.874**	0.654	0.747	0.881	0.512	0.647
Last_Name	929	**0.923**	**0.846**	0.883	**0.896**	0.791	0.839	0.868	0.612	0.718
Health_Care_Unit	1 025	0.744	**0.545**	0.629	0.531	0.351	0.422	**0.874**	0.281	0.425
Location	148	0.814	**0.375**	0.514	0.639	0.166	0.264	**0.872**	0.154	0.261
Phone_Number	137	0.824	**0.713**	0.764	0.569	0.181	0.275	**0.945**	0.561	0.704
Average	560	0.850	**0.712**	0.766	0.715	0.499	0.568	**0.891**	0.543	0.648

Table 1. Comparison of CRF++ with words and without words and random forests without words

for only two class labels. It should be noted that for de-identification purposes, we are normally most interested in reaching a high recall, something on which CRF++ clearly outperforms the two non-word approaches.

5. Conclusions and Future Work

It was argued that in order to allow for new methods and techniques to be developed and evaluated on real world clinical data that contain sensitive information, one option would be to provide access to derivations of such corpora without words (tokens), which instead are represented by sets of features that do not allow for any sensitive information to be derived. A requirement would then be that such non-word corpora should still contain relevant information. In this study, we investigated the effect on prediction performance when removing the actual words in a de-identification experiment using the Stockholm EPR PHI Corpus. It was observed that conditional random fields with access to the actual words clearly outperformed the same learning method, having access only to feature representations of the words, as well as random forests also considering only the latter features. The main conclusion is that the chosen set of features is not sufficient for representing the relevant information in this case, but additional features are needed in order to reach satisfactory performance. Such features may include more detailed annotations of where in the corpus the words are present, however the current feature rich and annotated clinical corpora can be released without the sensitive words for researchers that are interested in experimenting on finding better machine learning methods.

In the future work except of trying out a different feature set we would also try to use words as features in random forests to compare our results without using words. Another possibility is to keep e.g. function words in the corpus and give access to them since they are not sensitive. Yet another possibility is to use active learning to extend the annotated set and consequently the training set and then find a suitable feature set.

6. Acknowledgements

We would like to thank Elin Carlsson for excellent programming for the feature extraction. This work was partly supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by the Swedish Foundation for Strategic Research under grant IIS11-0053.

7. References

Alfalahi, A. (2011). Pseudonymization of person names in an annotated clinical Swedish corpus, Master thesis, Dept. of Computer and Systems Sciences, KTH/ Stockholm University. [https://daisy.dsv.su.se/fil/visa?id=62734]

Alfalahi, A., Brissman, S., and Dalianis H. (2012). Pseudonymisation of person names and other PHIs in an annotated clinical Swedish corpus. *Proceedings of The Third Workshop on Building and Evaluating*

Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, 26 May, Istanbul.

Boström, H. (2011). Concurrent learning of large-scale random forests. *Proceedings of Scandinavian Conference on Artificial Intelligence*, pp. 20–29.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L., (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L, Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, pp. 161–168.

CRF++ (2011). [http://crfpp.sourceforge.net/]

Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish clinical text – refinement of a gold standard and experiments with conditional random fields. *Journal of Biomedical Semantics*, 1:6 (12 April 2010)

Finkel, J.R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), pp. 363–370.

HIPAA (2003). *Health Insurance Portability and Accountability (HIPAA), Privacy Rule and Public Health Guidance, from CDC and the U.S.* Department of Health and Human Services. [http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm]

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, vol. 14, pp. 1137–1145.

Kokkinakis D. and Thurin A. (2007). Identification of entity references in hospital discharge letters. *Proceedings of 16th Nordic Conference on Computational Linguistics*, NODALIDA-2007, University of Tartu, Tartu.

Lafferty, J., McCallum A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings. 18th International Conference on Machine Learning*. Morgan Kaufmann, pp. 282–289.

Meystre, S.M., Friedlin, F.J., South B.R., Shen S., and Samor M.H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10:70.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization, the Regents of the University of California, *UCLA Law Review*, 57:1701–1819.

Olsson, F. (2008). Bootstrapping named entity annotation by means of active machine learning: a method for creating corpora. Doctoral thesis, University of Gothenburg.

Velupillai, S., Dalianis H., Hassel M., and Nilsson G. H., (2009). Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, doi:10.1016/j.ijmedinf.2009.04.005

Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus

Alyaa Alfalahi, Sara Brissman, Hercules Dalianis
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, 164 40 Kista, Sweden
E-mail: alyalfa@dsv.su.se, sarabri@dsv.su.se, hercules@dsv.su.se

Abstract

Today a large number of patient records are produced and these records contain valuable information, often in free text, about the medical treatment of patients. Since these records contain information that can reveal the identity of patients, known as protected health information (PHI), the records cannot easily be made available for the research community. In this research we have used a PHI annotated clinical corpora, written in Swedish, that we have pseudonymised. Pseudonymisation means to replace the sensitive information with fictive information for example real personal names are replaced with fictive personal names based on the gender of the real names and family relations. We have evaluated our results and our five respondents of who three were clinicians found that the clinical text looks real and is readable. We have also added pseudonymisation for telephone numbers, locations, health care units, dates and ages. In this paper we also present the entire de-identification and pseudonymisation process of a sample clinical text.

Keywords: Protected Health Information PHI, Electronic Patient Records EPRs, De-identification, Pseudonym, Swedish.

1. Introduction

Electronic patient records, EPRs, include valuable information about the treatment of the patient. Patient records also often contain sensitive information regarding the situation of the patient which may disclose private information about the patient; i.e. protected health information (PHI) such as names, locations, health care units, phone numbers, etc. (HIPAA 2003). Patient records include information about the patient such as their social situation, health history, symptoms, and previous diagnoses and planned treatment. Most of this information is presented in the unstructured free text (Dalianis et al, 2009) which can be extremely useful for the clinical researcher, for hospital management and for educational purposes. In Sweden, all research that deals with patients and data about patients requires permission for use from ethics committees, specifically regional Ethics Committees (Lag, 2003:460). The assumption is therefore that sensitive information must be removed from the records before EPRs can be made freely available for research, so the task of the de-identifying PHI instances is both important and difficult.

In Meystre et al. (2010) there is an overview of the different de-identification approaches for EPRs written in English. Most of the researchers have applied de-identification methods by identifying PHI instances and annotating them with PHI classes, e.g *First_Name, Last_Name,Health_Care_Unit,Location, Phone_Number,* etc. However, the records produced with these annotation classes are not easy to read as plain text. Similarly, the output text will be less readable if the PHI instances, such as personal names are replaced with ID numbers. As well as first and last names, there also phone numbers, locations, health care units, dates and ages that need to be pseudonymised and made readable. Some questions that arise involve how to replace the personal names and make the text coherent with respect to, for example, gender or

family relations, phone numbers that are realistic, locations that are geographically correct but not real, replacing ages without making patients too old or too young and finally changing times so they are realistic to weekends, seasons and public holidays. In this paper we will focus on first and last names.

2. Related Research

Meystre et al. (2010) reviewed different de-identification approaches for EPRs written in English, but they did not mention pseudonymisation. Pseudonymisation algorithms have, however, been mentioned in Sweeney (1996), Douglass et al. (2004), Pestian et al. (2005) and Neamatullah et al. (2008), who have all worked with EPRs written in English. Furthermore, Pantazos et al. (2011) focused on a de-identification algorithm for a Danish database and generated a new version by replacing real data with other new data. However, their algorithm has not been developed to handle unknown, misspelled names, or names that can be used for both genders.

Generally speaking there are few clinical corpora available for research in English and Finnish respectively.

- I2B2[1] corpus, the Informatics for integrating biology and the bedside (i2b2 2008) centre has created a clinical English corpus which consists of approximately 1,000 notes that is available for researchers after signing an agreement.
- The CMC[2] is one of the clinical corpora which has been analysed by Pestian et al. (2007) and

[1] http://www.i2b2.org
[2] http://www.inf.u-szeged.hu/rgai/bioscope

has been made public for study purposes.

- The De-id[3] corpus, Neamatullah et al. (2008), have published a clinical corpus in English. The De-id corpus consists of 412,509 nursing notes and 1,934 discharge summaries and is publicly available.
- A clinical Finnish corpus contains 2800 sentences (17,000 tokens) of nursing notes which have been manually anonymised by removing or changing all name. Furthermore, this corpus has been developed and published by Haverinen et al. (2010).

In Velupillai et al. (2009) an annotation process for de-identification is described. The annotation was carried out by three annotators, (one junior, one senior computer scientist and one senior physician) and gave rise to three sets of annotated data. Dalianis & Velupillai (2010) created a consensus of the annotated data in Velupillai et al. (2009) described with the de-identification system for Swedish clinical texts. The consensus is called the Stockholm EPR PHI Corpus, and contains 100 patient records with an equal distribution of male and female patients. These records were compiled from five clinics: neurology, orthopaedic, oral and maxillofacial surgery, an infection clinic and a dietetics clinic. However the Stockholm EPR PHI Corpus has not yet been pseudonymised. Another completely different approach is described by Dalianis & Boström (2012), who suggest releasing the Stockholm EPR PHI Corpus for research into different de-identification methods, augmenting the corpora with a large feature set but without giving access to the actual words. A non-worded corpora could be used for a set of machine learning experiments but unfortunately the feature set needs to be further developed.

3. Method

Our approach is to replace the PHI instances with new realistic instances. We call this process pseudonymisation. This is replacing real PHI instances in the clinical text with pseudonyms, surrogates or what we call fictive names. An automated algorithm (Pseudonoma) will be developed to replace the annotated first/last names in clinical free text notes with fictive names depending on the gender of the patient and the referential structure. Moreover, the algorithm has also been expanded to handle unknown and misspelled annotated personal names by continually checking the name lists. The algorithm will be executed on training (development) data and test data respectively. Finally we will also add pseudonymisation for *Phone_Number, Location, Health_Care_Unit, Full_Date, Part_Date* and *Age*.

The constructed pseudonymisation system (Pseudonoma) is a rule-based system with name lists. Pseudonoma is implemented partly in the Perl programming language, and consists of two algorithms, first name and last name algorithms, to replace real names with other fictive names. 'Name' could refer to patient's name, names of patient's relatives and health staff names. The other part of the Pseudonoma, for phone numbers, locations, health care units, dates and ages, was developed in Python and Excel script language. Pseudonoma has been tested on the Stockholm EPR PHI Corpus[4] (380,000 tokens) which is a subset of the Stockholm EPR Corpus (Dalianis et al, 2009).

The Stockholm EPR PHI Corpus is our input file to the first name algorithm, which includes annotated names with tags, such as *<First_Name> </First_Name>* and *<Last_Name> </Last_Name>*. The name lists have been created by retrieving personal names from Swedish names lists (Swedish names 2009) and have been manually checked regarding the gender of names. These name lists consist of a list of female first names (173 names), a list of male first names (114 names), and a list of last names (368 names). The programme contains several hash tables that store the processed names, inspired by Sweeney (1996). The output file from the first name algorithm is the input file to the last name algorithm. The final output file is our pseudonymised text.

We observed that annotated personal names in genitive form, misspelled or unknown personal names could cause problemd for our pseudonymisation algorithm and therefore negatively influence the readability and consistency of the pseudonymised text. Therefore, the algorithm (Pseudonoma) was improved to solve these problems. The algorithm was developed to handle misspelled names through the usage of edit distance (Levenshtein Distance [5]). The algorithm was also adjusted to deal with the genitive form and with typical Swedish name combinations such as *Anna-Lena, Eva-Britt, etc.* Also, unknown and gender- neutral names have been replaced with other gender- neutral names such as *Kim* and *Denis,* for details see (Alfalahi 2011).

4. Evaluation and Result

To evaluate Pseudonoma we have used two corpora. Firstly, the Stockholm EPR PHI Corpus, which consists of 100 patient records, has been used as training (or development) data for Pseudonoma. The 100 patient records have previously been manually annotated (Velupillai et al. 2009). Secondly, a new extract from the Stockholm EPR Corpus that also includes 100 patient records has been used as test data which has been annotated automatically by applying the Stanford CRF NER programme (Dalianis & Velupillai 2010). We have executed the Pseudomona on both above mentioned corpora. In Figure 1 we can see the complete automated

[3] http://www.physionet.org/physiotools/deid

[4] The study was carried out after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

[5] http://en.wikibooks.org/wiki/Algorithm_implementation/Strings/Levenshtein_distance#Perl

de-identification and pseudonymisation process: firstly the clinical text annotated with respect to PHI by the Stanford CRF NER, trained on the Stockholm EPR PHI Corpus, and then the annotated text pseudonymised by Pseudonoma.

The outcome is pseudonymised patient records that contain fictive names. 14 records were chosen from the Stockholm EPR PHI Corpus and 12 records were selected from a new extract from the Stockholm EPR Corpus for manual evaluation by five respondents, three of whom were clinicians. In Figure 2 we can see the distribution of names and tokens in our data.

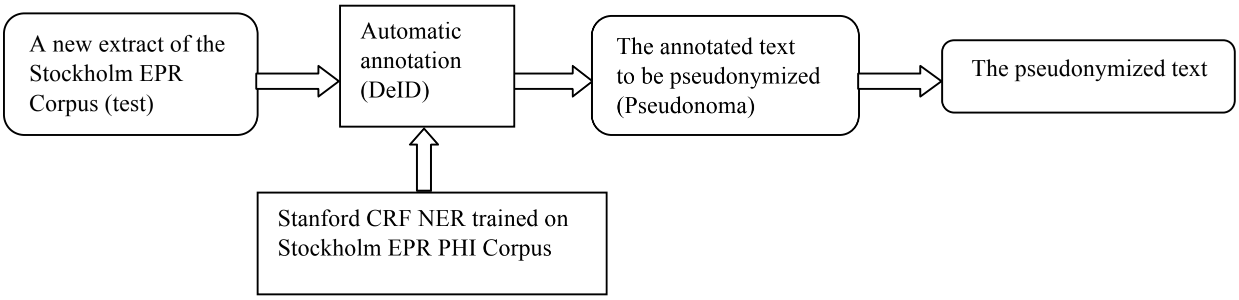

A new extract of the Stockholm EPR Corpus (test)
Automatic annotation (DeID)
The annotated text to be pseudonymized (Pseudonoma)
The pseudonymized text
Stanford CRF NER trained on Stockholm EPR PHI Corpus

Figure 1: The complete de-identification and pseudonymisation process using Stanford NER CRF and Pseudonoma

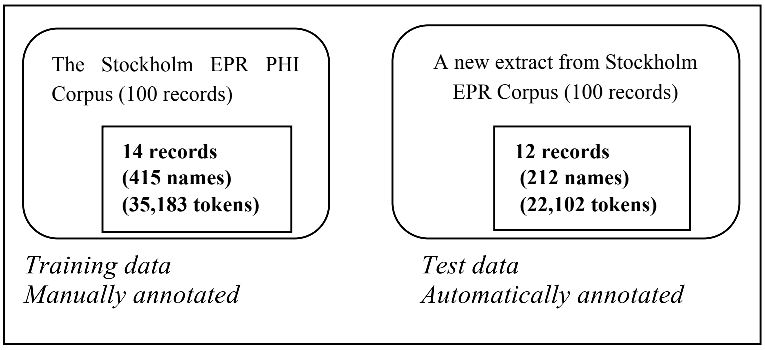


Figure 2: Comparison of the two evaluation sets and their distribution of names

Questions	Total names	Yes (%)	No (%)
Gender determination[6]	97	95 (98)	2 (2)
Gender agreement[7]	97	97 (100)	0 (0)
Difference between first and last name	212	212 (100)	0 (0)
Pseudonymisation	212	212 (100)	0 (0)
Repetition of the same fictive name	212	212 (100)	0 (0)
First/ last name tags are replaced	212	212 (100)	0 (0)

Table 3: Answers to the questionnaire (questions 1-6) on test data (212 first and last names, 97 first names)

[6] The fictive names have the right gender according to the text.

[7] The fictive name has the same gender as the real name, i.e. if the real name has a female gender so a female name must be chosen as a fictive name.

We created a questionnaire with 11 questions to be answered by each respondent. Two types of results were obtained by applying the questionnaire to the patient records (the original and the pseudonymised records): the results according to the number of records in both corpora (training and test data) and the results according to the number of annotated names in both corpora. Table 3 illustrates the responses to six questions depending on the number of annotated names in the records.

The aim of our questionnaire was to evaluate six different values that reflect the quality of the text. The *gender determination* of the fictive names means that the fictive names have a clear gender so it is easy to distinguish between female and male fictive names in the output text. G*ender agreement* means that if the real name is female in gender (Eva) so a female name must be chosen as the fictive name (Sara). The *difference between first and last name* means that the respondent can specify the first and the last names from the correct selection of fictive names such as *Johan* for first name and *Johansson* for last name. Another evaluation point that has been added to the questionnaire concerns the replacement (*pseudonymisation*) of the real name with the fictive and whether all real names have been replaced with fictive names. A further evaluation point in the questionnaire relates to the *repetition of the same fictive name* in the text, specifically, whether the real name has been replaced with the same fictive name each time the real name appears in the text. For example, the real name *Erik* should have the same fictive name *Tomas* whenever *Erik* is repeated in the text. The last question concerns left over tags *(<First_Name> </First_Name>,<Last_Name> </Last_Name>*) in the patient records.

Our goal was to develop an automated algorithm which can correctly replace all annotated real names with other real names (100 per cent) in patient records. The algorithm developed was tested on the Stockholm EPR PHI Corpus (the development or training data) and on a new extract from Stockholm EPR Corpus as test data.

The questionnaire analysis shows that the main goal is achieved by correctly replacing all annotated names with other realistic names. The automated algorithm depends exclusively on the annotation process. There were two un-annotated names in the chosen records (14 records, 415 names) from the Stockholm EPR PHI Corpus (training data) and 16 un-annotated names in the Stockholm EPR Corpus (test data) (12 records, 212 names), and so these un-annotated names were not replaced by the algorithm. This made the training text slightly more readable and coherent than the test text.

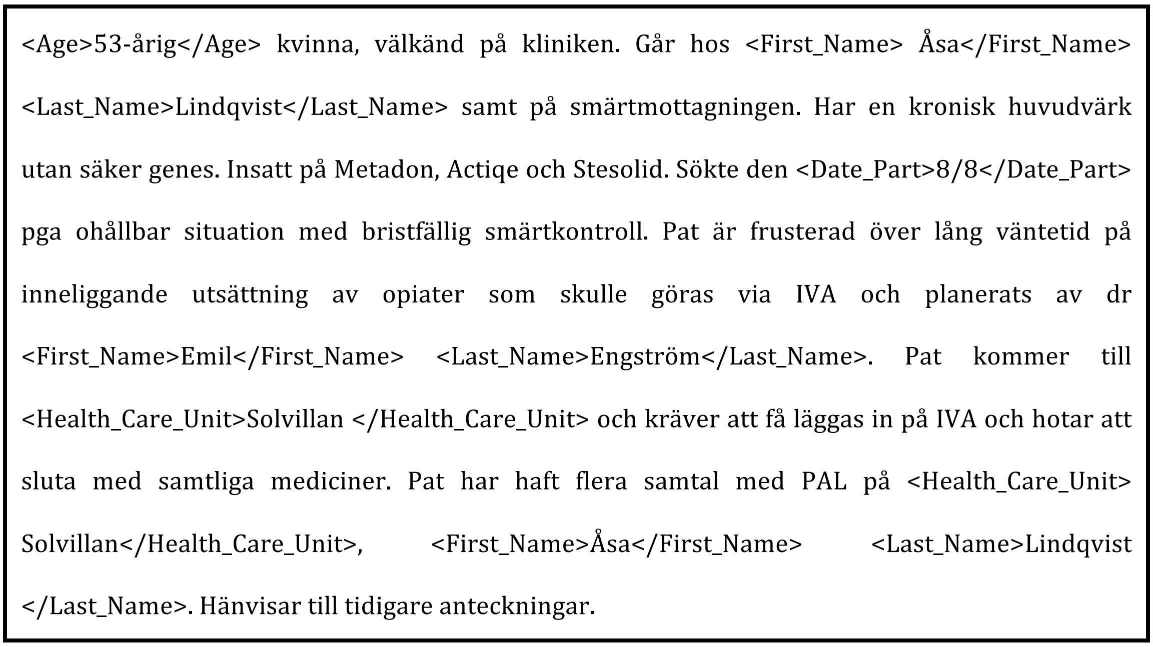
The questionnaire analysis illustrates that selection of the right gender during the replacement process is achieved to a high percentage in training and in the test corpora, 99 and 98 per cent respectively. Furthermore, the genitive form, and misspellings have also obtained a high percentage of accuracy in both corpora. The question about genitive form includes a check of the genitive *s* in the fictive name if the real name takes the form of genitive *s*. For example, if the real name is in the genitive form i.e. *Eva's* mother (Evas mor), then the fictive name for *Eva*, i.e. (*Karin*) must definitively have the genitive *s*, *Karin's* mother (Karins mor). Another question tests whether the misspelling of real names has been handled by the correction technique, which has been improved in the algorithm. Additionally, the processing of Swedish characters (*äåö ÄÅÖ*) is not standard in Perl language so the algorithm has been developed to handle these types of characters which can occur in names such as *Märta*, *Håkan*, *Göran*, Åsa, etc. This handling of language-specific problems obtained a high percentage of accuracy in both corpora.

We continued the pseudonymisation work by pseudonymising Phone*_Number, Location, Health_Care_Unit, Full_Date, Part_Date* and *Age*. Altogether we pseudonymised 4421 instances in our corpus distributed over the classes described in Table 4.

The second part of Pseudonoma was developed in Python and Excel scripts except for one section regarding ages, which was manual as the number of ages were few. All dates were shifted by an unknown, arbitrary number of days and months respectively. *Phone_Number* was pseudonymised except for the area code and finally *Location* and *Health_Care_Unit* was assigned to the default location Stockholm and default health care unit Solvillan respectively, which was a naive approach. Age was manually shifted by an unknown arbitrary number of years except for *Age over 89* years which was shifted to Age over 89. Please see Figure 5 for an example of this.

Annotation class	Instances
Age	56
Full_Date	710
Date_Part	500
First_Name	923
Last_Name	928
Location	1 021
Health_Care_Unit	148
Phone_Number	135
Sum	4 421

Table 4: The distribution of annotation classes and instances of pseudonymised annotation



<Age>53-årig</Age> kvinna, välkänd på kliniken. Går hos <First_Name> Åsa</First_Name> <Last_Name>Lindqvist</Last_Name> samt på smärtmottagningen. Har en kronisk huvudvärk utan säker genes. Insatt på Metadon, Actiqe och Stesolid. Sökte den <Date_Part>8/8</Date_Part> pga ohållbar situation med bristfällig smärtkontroll. Pat är frusterad över lång väntetid på inneliggande utsättning av opiater som skulle göras via IVA och planerats av dr <First_Name>Emil</First_Name> <Last_Name>Engström</Last_Name>. Pat kommer till <Health_Care_Unit>Solvillan </Health_Care_Unit> och kräver att få läggas in på IVA och hotar att sluta med samtliga mediciner. Pat har haft flera samtal med PAL på <Health_Care_Unit> Solvillan</Health_Care_Unit>, <First_Name>Åsa</First_Name> <Last_Name>Lindqvist </Last_Name>. Hänvisar till tidigare anteckningar.

Figure 5. Example of pseudonymised Swedish clinical text by Pseudonoma, where all annotated instances have been replaced by pseudonymised instances. (In the real output text the annotation tags are, of course, removed).

All our data which could reveal a patient's identity, such as corpora, name lists and hash tables, is stored encrypted.

5. Conclusions and Future Work

The main contribution of this paper is that pseudonymisation, i.e. replacement of real names with fictive names in electronic patient records written in Swedish with an automated algorithm, is possible to a high quality (100%). Maintenance of the patient's gender and family relationships makes the text readable and coherent to a high quality (100%). The process of pseudonymisation is exclusively dependent on the quality of annotation of the text, whether manual or automatic. To the best of our knowledge this is the first algorithm that has been developed to automatically replace all real names with other real names in Swedish clinical text, as well as for phone numbers, location, health care units, dates and ages. The pseudonymised text may additionally be used for medical educational purposes or the development of new tools, and it is therefore important that the text maintains the readability. We believe the algorithm for pseudonymisation can be easily adapted to other languages, one only need change the name lists to the local language.

We have also in this paper showed the complete de-identification process of a sample clinical text (a new extract) using a machine learning system Stanford NER CRF trained on the manually annotated Stockholm EPR PHI Corpora to de-identify (annotate the PHI) the sample text, followed by the pseudonymisation of the sample clinical text using Pseudonoma, (see Figure 1).

In the future we would like to extend the location and health care unit pseudonymisation to select similar general locations and health care units to those written in the patient record. One issue that arises is that if one replaces health care units one may miss important information about diseases. Date shifts are also sensitive, cannot be completely randomized and have to be consistent; one needs to consider that weekends, public holidays and the seasons are different in respect of health care, as during weekends and public holidays there are fewer health care personnel on duty and some diseases are seasonally dependent.

In the future we plan to add automatic age replacement, and to evaluate the performance and quality of the pseudonymisation of phone numbers, locations, health care units and ages. We also plan to manually re-read the entire corpus one more time to be sure that we have not missed annotations for any PHI or pseudonymising any PHI. We have also previously tested Stanford NER CRF trained on the same corpus and found another 49 false positives that have been annotated (Dalianis & Velupillai 2010).

We are currently in the process of applying for ethical permission to release this pseudonymised variant of the Stockholm EPR PHI Corpus, which we call the Stockholm EPR PHI Pseudo Corpus.

6. Acknowledgements

This work was partly supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by the Swedish Foundation for Strategic Research under grant IIS11-0053.

7. References

Alfalahi, A., (2011). Pseudonymization of person names in an annotated clinical Swedish corpus, Master thesis, Department of Computer and Systems Sciences (DSV) KTH/Stockholm University. Internet: https://daisy.dsv.su.se/fil/visa?id=62734

Dalianis H. and Boström, H. (2012). Releasing a Swedish clinical corpus after removing all words - de-identification experiments with conditional random fields and random forests, in Proceedings of The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul.

Dalianis, H., Hassel, M. and Velupillai, S. (2009). The Stockholm EPR Corpus - Characteristics and some initial findings, Proceedings of ISHIMR (2009), Evaluation and implementation of e-health and health information initiatives: International perspectives, 14th International Symposium for Health Information Management Research. Kalmar, Sweden, 14-16 October, pp. 243-249.

Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields, Journal of Biomedical Semantics, 1:6 (12 April 2010).

Douglass, M., Clifford, G., Reisner, A., Moody, G. and Mark, R. (2004). Computer-assisted de-identification of free text in the MIMIC II database, Computers in Cardiology 31: 341–344. Internet: http://mimic.mit.edu/Archive/Publications/Douglass04.pdf.

Haverinen, K., Ginter, F., Laippala, V., Viljanen, T. and Salakoski, T., (2010). Dependency- based PropBanking of clinical Finnish, In Proceedings of The Fourth Linguistic Annotation Workshop (LAW IV) held at ACL2010, Uppsala, Sweden. Internet: http://bionlp.utu.fi/clinicalcorpus.html

HIPAA (2003). Health insurance portability and accountability (HIPAA), privacy rule and public health guidance, From CDC and the U.S. Department of Health and Human Services, http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm.

I2b2, (2008) Informatics for integrating biology and the bedside, Internet: http://www.i2b2.org.

Lag (2003:460) om etikprövning av forskning som avser människor (SFS). Stockholm: Utbildningsdepartementet. (In Swedish, Law (2003:460) Ethical review regarding research considering humans), Internet: http://www.notisum.se/rnp/sls/lag/20030460.HTM

Meystre, S., Friedlin, F., South, B., Shen, S. and Samore, M. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Medical Research Methodology, 10:70.

Neamatullah, I., Douglass, M., Lehman, L., Reisner, A., Villarroel, M., Long, W., Szolovits, P., Moody, G., Mark, R. and. Clifford, G. (2008). Automated de-identification of free text medical records, BMC Medical Informatics and Decision Making, 8: 32. Internet: http://www.physionet.org/physiotools/deid

Pantazos, K., Lauesen. S. and Lippert, S. (2011). De-identifying an EHR Database – Anonymity, Correctness and Readability of the Medical Record, European Federation for Medical Informatics.

Pestian, J. P., Itert L., Andersen C. L. and Duch W. (2005). Preparing clinical text for use in biomedical research, Journal of Database Management, 17(2):1-12.

Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K. and Duch, W. (2007). A shared task involving multi-label classification of clinical rree text, BioNLP : Biological, translational, and clinical language processing, pages 113–120. Prague, June, Association for Computational Linguistics. Internet: http://www.inf.u-szeged.hu/rgai/bioscope

Swedish names, in Swedish (2009). Internet http://svenskanamn.se

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system, Proceedings, Journal of the American Medical Informatics Association, pp. 333-337.

Velupillai, S., Dalianis, H., Hassel, M. and Nilsson, G. (2009). Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial, International Journal of Medical Informatics, 78, e19-e26.

Evaluation and Performance Improvement of the BioExcom System for the Automatic Detection of Speculation in Biomedical Texts

Olfa Makkaoui, Julien Desclés, Jean-Pierre Desclés
LaLIC-STIH
Université de Paris-Sorbonne
Maison de la Recherche
28 rue Serpente, 75006 Paris
E-mail: olfa_makkaoui@yahoo.fr, julien.descles@gmail.com, jean-pierre.descles@paris.sorbonne.fr

Abstract

The BioExcom system aims to automatically annotate speculative sentences in biomedical texts and to categorize them into "*new*" and "*prior*" speculations. This work highlights a more restrictive way to consider speculations as a source of knowledge for biologists who are also interested in finding hypotheses in the biomedical literature. The system is based on the Contextual Exploration processing (hierarchical research of linguistic surface markers with the EXCOM computational platform). The BioExcom evaluation is realized on the BioScope corpus by manually comparing the BioExcom automatic annotations and the BioScope manual annotations. The analysis of diverging annotations was a starting point to build a new version of the system (BioExcom_2) that results from the performance improvement of the initial system BioExcom. A corpus *BioSpe* for the annotation of speculative sentences is established. This latter was annotated according the BioExcom characterization of speculation and can be used by machine learning systems. A user interface for the automatic annotation of speculative sentences is made available on line.

Keywords: speculation, hypothesis, biology, automatic annotation, text mining

1. Introduction

Recent research in text mining linked to the biological domain has made major progress and took into consideration the importance of extracting speculation by distinguishing between factual statements and uncertainty (Medlock 2007, Kilicoglu and Bergler, 2008). This task is especially linked to the consideration that Biological researchers can be only interested in finding factual sentences in the text. Information is consequently classified as certain or speculative. These latter are considered in this case as hedges since their meaning concerns all information that do not belong to the certain statements.

However, biologists can be also interested in extracting speculations linked for example to a particular entity (Light et al., 2004). This task is important for their experimental research as authors are not sure about their results and speculations they provide can be a starting point for new experiments (De Waard, 2009). The SWAN project illustrates an example of the usefulness of such statements as it aims to collect hypothetical information about the Alzheimer disease in order to use it as discussion subjects between researchers (Ciccarese et al., 2008). The meaning of speculation is in this case more restrictive than hedges and is very close to hypothetical statements.

This latter speculation characterization is developed by the BioExcom system that aims to answer to the biologists needs concerning the speculation extraction in biological texts (Desclés et al., 2009). This approach underlines the importance of establishing a link between their experimental findings and ideas or proposals about biological issues provided in the literature without taking into account approvals or negations of them.

The BioExcom system categorizes also speculation into "*new*" and "*prior*" speculations.

The annotation methodology is based on the Contextual Exploration processing developed in the EXCOM engine that requires the use of linguistic resources linked to a particular semantic category (speculation in the case of the BioExcom system).

The BioExcom system performance was evaluated on a small corpus (Desclés et al., 2009). In this paper, we aim to evaluate BioExcom on a large scale using an independent corpus like BioScope and also to compare the two characterizations of speculation and see in what they differ (Szarvas et al., 2008). This evaluation step consists in automatically annotating a part of the BioScope corpus (14500 sentences) using BioExcom and then manually comparing the results (BioScope manual annotations and BioExcom automatic annotations). This processing reveals an important number of converging annotations. In order to evaluate correctly the BioExcom system, diverging sentences are manually analyzed with the consideration of two hypothesis stating that converging sentences are correctly classified (as speculative or not speculative sentences).

We aim in this paper:

- To improve the BioExcom system performance basing on the comparison between the BioExcom automatic annotations and the BioScope manual annotations.
- To present a new copus *BioSpe* for the annotation of speculative sentences according to the more restrictive characterization of speculation.
- To present an online user interface that enables the

automatic annotation of speculative sentences and their categorization into "*new*" and "*prior*" subcategories.

2. Related Work

Hyland (1998) proposes a description of hedging[1] in scientific articles by presenting a pragmatic classification of hedge cues resulting from the annotation of a corpus of molecular biology articles where hedge cues are classified as model auxiliaries, epistemic lexical verbs, epistemic adjectives, adverbs and nouns. According to the author, hedging in scientific articles can be used to weaken statements or signal uncertainty.

In (Friedman et al., 1994) clinical information in patient documents are translated into controlled vocabulary using semantic grammar based rules. Information is classified according to five certainty types namely no certainty, low certainty, moderate certainty and cannot evaluate.

Light et al., (2004) focus on extracting expressions of belief by manually classifying sentences as definite, high speculative and low speculative arguing that the low speculative level is used to express a statement following almost directly from results but not quite whereas high speculative statements contain a more dramatic leap from the results. This study concluded that it is not possible to distinguish between the two statements. A Support Vector Machine classifier was also used to automatically classify abstract sentences as speculative or definite.

This work was extended by Medlock and Briscoe (2007) and proposes detailed definition of hedge by providing an annotation guideline. A weakly supervised machine-learning model using SVM is performed to classify sentences as speculative or non-speculative using other features like part of speech, lemmas and bigrams.

Szarvas (2008) aims to classify sentences as speculative or non-speculative in radiology reports and scientific articles using a weakly supervised machine where feature consists of a selection of word extracted either manually or automatically. Additional keywords extracted from external dictionaries are also used to improve the classification performance.

The system was evaluated on a data set gathered and made available by Medlock and Briscoe (2007) and obtained an F-Measure of 85% on the Fly Base data set and 75% of BMC bioinformatics data set.

Kilicoglu and Bergler (2008) use knowledge from existing linguistic and lexical resources and incorporate pattern to build a classifier that enables the speculative sentence recognition. The system was tested on new test sets: The first one consists of a corpus made publicly available by Medlock and Briscoe (2007) whereas the second data set was provided by Szarvas et al., (2008).

In (Morante and Dealmans, 2009) hedging and their scopes are detected based on a two stage classification task. A set of classifier is used to identify hedge cues then the scopes are detected by another set of classifier in the second stage.

Agarwal (2010) focused on detecting hedging and their cues in biomedical literature using a supervised algorithm trained on the BioScope corpus (Szarvas et al., 2008) and obtain an F1 score of 88% and 86% in detecting hedge cues and their scope in biological literature and an F1 score of 93% and 90% in detecting hedge.

3. The BioExcom System

According to the BioExcom characterization, speculation in the biomedical literature is a proposal about a biological issue that is explicitly presented as not certain in the paper. This information can deal with working hypothesis, possible interpretations or explanations of a fact or purely speculative statements (theoretical considerations). This implies that statements such as deductions, conclusions or demonstration are not considered as speculative.

The BioExcom system aims to annotate speculative sentences and to categorize them into "*new*" and "*prior*" speculations. To detect speculative sentences, BioExcom uses the Contextual Exploration processing (Desclés et al., 2006) that is based on the search for linguistic markers (indicators and clues) presented by regular expressions to annotate textual segments (which can be a title, a paragraph or a clause) depending on a given discursive category (definition, result, speculation…).

As the simple detection of these indicators are, in some cases, not sufficient to correctly annotate sentences, the Contextual Exploration processing focuses on some other linguistic markers (clues) in the indicator context to remove ambiguities. Linguistic clues can be positive if they enable to confirm an annotation decision or negative if they are used to invalidate it. This process is useful to resolve some ambiguous linguistic markers such as the "*remains unknown*" indicator of the sentences (1) and (2). Although both of them use the same indicator, they express two different meanings. Indeed, the presence of the "*whethe*r" clue indicates that the sentence (1) is a speculation whereas the "*how*" clue shows that the sentence (2) expresses a lack of knowledge. This latter notion deals with open questions without presenting any proposal or idea about a subject.

(1) "Also, whether the signaling activity of Ser is similarly regulated by endocytosis remains unknown".

(*2) "How endocytosis of DI leads to the activation of N remains unknown."*

Figure1 illustrates the successive steps for the automatic semantic annotation:

- Step1: Looking for indicators in the segment
- Step 2: Call and execution of the associated contextual rules which are triggered by the identification of an indicator in the sentence.
-Step 3: Looking for clues contained in the rule. This search can be performed in the sentence at the right

[1] The term hedging was introduced by Lakoff (1972) to describe absence of certainty and is employed to indicate either a lack of commitment to the truth value of an accompagnying proposition or a desire not to express that commitment categorically.

or/and at the left of the indicator or even inside the indicator.
-Step 4: Semantic annotation of the segment if all the rules conditions are satisfied.

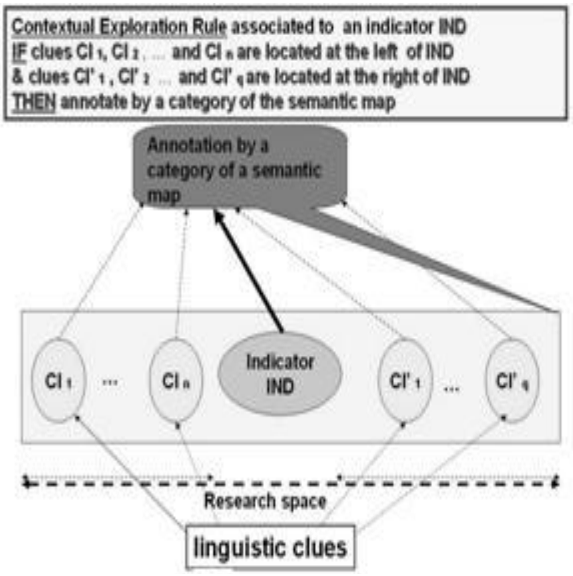

Contextual Exploration Rule associated to an indicator IND
IF clues Cl1, Cl2, ... and Cln are located at the left of IND
& clues Cl'1, Cl'2 ... and Cl'q are located at the right of IND
THEN annotate by a category of the semantic map
Annotation by a category of a semantic map
Cl1
...
Cln
Indicator IND
Cl'1
...
Cl'q
Research space
linguistic clues

Figure 1: The contextual exploration processing: search for an indicator and then for some clues in a contextual space (a sentence or a clause in our case) according to some associated rules (IND is indicator and CL1...CLn are clues.

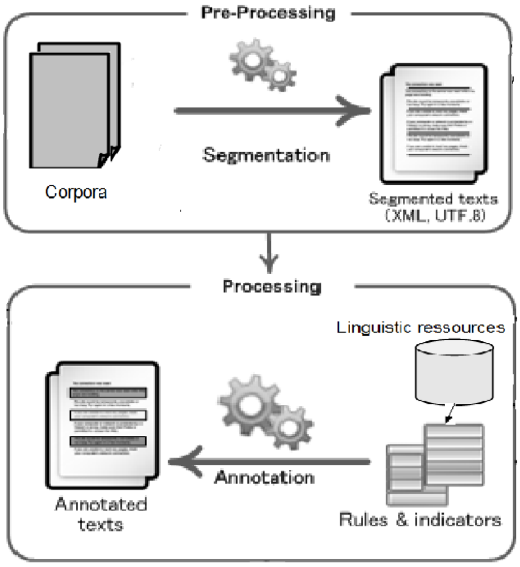


Figure 2: Overview of Excom processing (Alrahabi and Desclés, 2009)

The Contextual Exploration Processing is performed by the EXCOM platform (Djioua et al., 2006) (Alrahabi and Desclés, 2009) that enables to automatically annotate texts according to a given view point (definition, results...). The system general architecture is described in Figure 2. Texts are segmented using a list of typographical signs. The obtained segments (which can be sections, paragraphs or sentences) are then automatically annotated using the Contextual Exploration processing.
In BioExcom the speculation linguistic markers extraction and the Contextual Exploration rules construction were carried out by a biologist and a linguist on about seventy biological texts. The BioExcom annotation process requires thirty Contextual Exploration rules based on twenty semantic and grammatical indicator categories (Desclés et al., 2009).
The detection of speculative sentences requires the use of different linguistic types such as nouns, modality verbs, adverbs and conjunction (Desclés et al., 2009).
In some cases, the simple presence of some markers was sufficient to annotate a sentence as a speculation such as *"may"* modality verb in the sentence (3). Other indicators require a context analysis by looking for additional clues in the sentence to validate or not the annotation decision such as the "*remains unknown*" indicator in the sentence (1).

(3) "*Solute transport by GmNod26 may be related to a role in osmoregulation of the peribacteroid space".*

The categorization of speculative sentences into *"new"* and "*prior"* subcategories task was based on the search for some specific verbal aspects and also specific linguistic clues. Indeed, to annotate a sentence as a "*new speculation*", BioExcom looks for the absence of bibliographic citation or the presence of specific words such as *"in this study"* in the sentence.
The annotation of a sentence as "*prior speculation*" depends on the presence of bibliographic citation and some specific words like *"recent report"* as positive clues.

4. BioExcom Evaluation

The speculation detection task was first evaluated on a small corpus and enabled to prove the method's effectiveness (Desclés et al., 2009). A following step is realized in this study in order to evaluate the BioExcom performance on a large scale concerning the detection of speculative sentences using a new corpus like BioScope (Szarvas et al., 2008). It consists of three parts namely medical free texts, biological full papers and biological scientific abstracts. Only the biological full papers and the biological scientific abstracts parts (consisting of 9 full-texts and 1273 abstracts) of the BioScope corpus were analyzed because the BioExcom system is especially interested in analyzing the biomedical scientific domain.

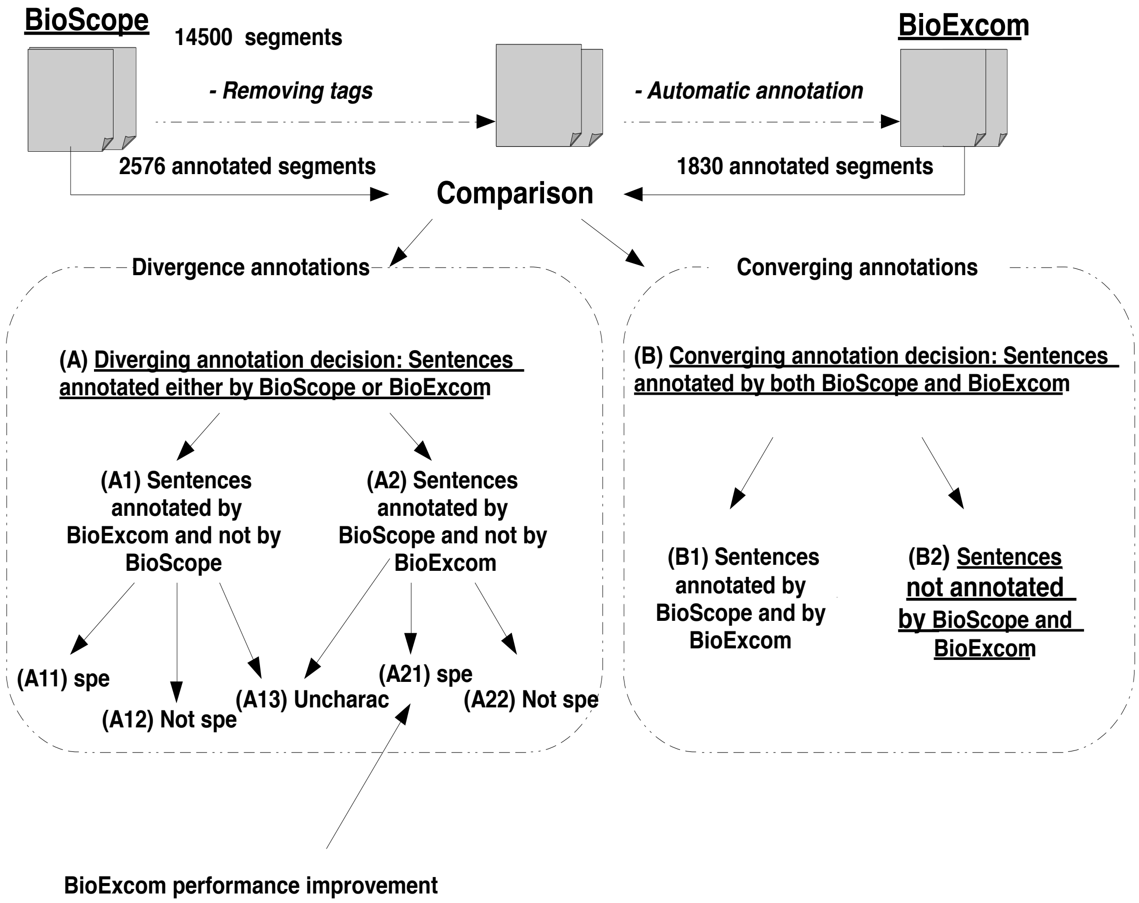

BioScope
14500 segments
BioExcom
- Removing tags
- Automatic annotation
2576 annotated segments
1830 annotated segments
Comparison
Divergence annotations
Converging annotations
(A) Diverging annotation decision: Sentences annotated either by BioScope or BioExcom
(B) Converging annotation decision: Sentences annotated by both BioScope and BioExcom
(A1) Sentences annotated by BioExcom and not by BioScope
(A2) Sentences annotated by BioScope and not by BioExcom
(B1) Sentences annotated by BioScope and by BioExcom
(B2) Sentences not annotated by BioScope and BioExcom
(A11) spe
(A12) Not spe
(A13) Uncharac
(A21) spe
(A22) Not spe
BioExcom performance improvement

Figure 3: BioExcom large scale evaluation: comparison with the annotations of the BioScope corpus part spe=speculation, not spe=not speculation, Uncharac=Uncharacterized

	(A) Diverging annotations		**(B) Converging annotations**	
Manual Checking	(A1) Sentences annotated by BioExcom and not by BioScope	(A2) Sentences annotated by BioScope and not by BioExcom	**(B1)** Sentences annotated by BioExcom and BioScope	**(B2)** Sentences not annotated by BioExcom and BioScope
Spe	97	379		
No Spe	17	437		
Uncharacterized	5	49		
Total	119	865	1711	12 652

Table 1: Statistics of the manual checking of the BioScope and BioExcom annotation (annotation decision spe=speculation, not Spe=not speculation)

The BioScope two parts annotations tags were first removed then automatically segmented and annotated by the BioExcom system (steps are presented in Figure 3).

BioExom automatically annotated 1830 sentences (341 sentences from full text papers and 1489 sentences from the abstracts corpus part).

The evaluation results were calculated according to the BioExcom segmentation due to the presence of a few sentences that were not segmented exactly in the same way by BioExcom and BioScope. The categorization into *"new"* and *"prior"* speculation was not taken into consideration during the evaluation process.

Table1 illustrates the comparison results of the BioExcom and the BioScope annotations. The evaluation results are presented in table 2. The Precision is approximately 93% in average (calculated from the total of segments of the two corpora) and the Recall is approximately 68% (in average). Comapred to the BioExcom first evaluation (Desclés et al. 2009), the recall dramatically falls.

	Precision	Recall	F-Measure
Full Text Papers	89,35	62,92	73,84
Abstracts	94,75	68,83	79,74

Table 2: Summary of raw results for the BioExcom evaluation

In order to evaluate the performance of BioExcom according to its own definition of speculation and to analyze the observed low Recall (Table 2), annotated sentences were compared.

The annotations comparison reveals the presence of two sentences categories (converging and diverging sentences). The sentences analysis task was performed by a biologist and two linguists not allowed to communicate with each other and did not know whether the annotations were performed by BioScope or BioExcom. Conflicts[2] were resolved by discussions during regular meetings and, in case of important uncertainty for at least two annotators, the sentences, called uncharacterized (54 in total), were not taken into consideration.

Converging annotations (B) (evaluated on a sample of the BioScope corpus part):

The converging annotations (B sentences group) contain sentences annotated by both BioExcom and BioScope (the B1 sentences group) and also sentences annotated neither by BioExcom nor by BioScope (the B2 sentences group). The evaluation process is based on the hypotheses that BioExcom and BioScope, when converging (case of annotation or not annotation), took the right decision.

A first hypothesis states that all the sentences that belong to the B1 sentences group are speculative what means that BioExcom and BioScope did not wrongly consider a non-speculative sentence as a speculation. A second hypothesis states that the B2 sentences group does not contain speculative sentences (B2∩spe= Ø). This suggests that BioExcom and BioScope did not forget to annotate a same speculative sentence.

The validation of the BioExcom large scale evaluation requires the analysis of the two previously presented hypotheses. To realize this task, we chose to evaluate manually a random sample of 10% of the BioScope part corpus that was initially evaluated. This corpus sample consists of 2 full texts and 130 abstracts (1823 sentences). After removing the annotations tags, the sample was manually annotated by three evaluators that were not allowed to communicate with each other except to discuss ambiguous cases. Annotations agreement between the evaluators implies a validation of a decision (whether the decision concerns a sentence annotation or not).

The sample evaluation results are presented in table 3 and confirm the validation of the two hypotheses concerning the agreement between BioExcom and BioScope.

The sample evaluation reveals that sentences which were not considered as speculation by both BioExcom and BioScope) contain some speculative sentences. Indeed, BioExcom and BioScope forgot to annotate 16 speculative sentences (from a total of 265 speculative sentences found in the sample) dealing with the same linguistic marker.

For example, in the sentence (4), the indicator "*evidence that*" is not used as a linguistic marker by both BioExcom and BioScope to annotate speculation.

(4) *"Athough NF-AT has not been cloned or purified, there is evidence that it is a major target for immunosuppression by cyclosporin A (CsA) and FK506 (refs 2-7)"*.

	Sentences from the sample that were annotated by BioExcom and BioScope as speculation	Sentences from the sample that were not annotated by BioExcom and BioScope as speculation
Spe	265	16
Not Spe	2	1540
Total	267	1556

Table 3: Summary of the manual evaluation results of the corpus sample (Spe=speculation, Not spe=Not speculation)

Diverging annotations (A):

This category deals with sentences that belong to either BioExcom annotations (A1 sentences group) or BioScope annotations (A2 sentences group) but not by both of them. . The manual checking of these diverging sentences showed that an important number of not speculative sentences are annotated by BioScope (the A2 ∩ Not Spe sentences group). This step reveals some critical points concerning the BioScope corpus.

First, the sentences comparison test confirms that BioScope and BioExcom do not have the same speculation characterization. Indeed, according to the BioExcom speculation view, information in biological papers depends on different certainty level namely certain statements (results, data, observation…), uncertain statements (speculation) and intermediary statements (demonstration, deduction….). Sentences expressing deductions or demonstrations (intermediary statements according to BioExcom) are not considered as speculative while BioScope annotated some of them. For example, the sentence (5) and (6) are annotated by BioScope whereas the indicator "*can be deduced that*" in the sentences (5) and "*indicate that*" in the sentence (6) are, according to the BioExcom speculation characterization, used to express rather a deduction than a speculation. Indeed these sentences present things more or less as certain.

[2] The annotation guidelines is available on www.bio-excom.net/corpus.htm

This characterization is in agreement with Thompson et al., (2008) who also showed that these linguistic markers can be used to detect deductive statements and treated the speculative one in another Knowledge Type category. In this view, the case of "*indicate that*" is interesting to be detailed. Whereas many studies use it as a linguistic marker of speculation, Kilicoglu and Bergler (2008) moderated its speculative meaning by highlighting the additional need to take into account its context.

(5) "It can be deduced that the erythroid ALAS precursor protein has a molecular weight of 64.6 kd, and is similar in size to the previously isolated human."

(6) "These findings indicate that corticosteroid resistance in bronchial asthma cannot be explained by abnormalities in corticosteroid receptor characteristics".

Second, some wrong annotations are detected during the manual annotation process which is the case in sentences (7) and (8). Indeed, the indicator "*or*" present in the sentence (7) can be replaced by "*and*" which implies that the sentence rather deals with factual information than speculation.
In addition, in the sentence (8), the indicator "*could*" is used to express the past form of the verb "*can*" and not its conditional form.

(7) "*To perform such a comparison, the EOCT predictor (Expression, Orthology, Combined and Transitive modules) was trained on datasets consisting of either equal numbers of positives and negatives or 100 times more negatives than positives and then tested on both types of datasets*".

(8) "The c-erbA-dependent activation of this CAII reporter construct could only be suppressed by very high amounts of v-erbA."

Third, BioScope annotates sentences expressing "*lack of knowledge*" or "*open questions*" as speculation (case of sentences (9) and (10)) which is contradictory to the BioExcom annotation purpose since it considers that sentences dealing with "*lack of knowledge*" or *"open question*" do not provide any proposal.

(9) "Because point mutagenesis cannot distinguish between family members, it is not known which protein activates 5".

(10) "The mechanism by which progesterone causes localized suppression of the immune response during pregnancy has remained elusive"

Finally, BioScope did not annotate a group of speculative sentences. As an illustration, the following sentence is clearly a speculation ("*We hypothesize that*") but was not annotated in the BioScope corpus.

(11) "*We hypothesize that a mutation of the hGR glucocorticoid-binding domain is the cause of cortisol resistance".*

From this work, the BioSpe[3] corpus was established and made available on line[4]. It is based on the converging sentences between the BioExcom automatic annotations and the BioScope manual annotations (the correctness of these annotations has been checked on a sample presented in the analysis of converging annotations part) and the manual annotations of diverging sentences
To correctly evaluate the BioExcom system, we recalculate the precision, recall and F-Measure, according to the BioSpe corpus of speculations (results are illustrated in table 4). Corrected Precision, Recall and F-Measure are respectively around 99%, 83% and 90% (averages calculated from the total of segments of the two corpora).

	Precision	Recall	F-Measure
Full Text Papers	97,63	77,46	86,39
Abstracts	99,39	83,93	91,01

Table 4: BioExcom evaluation based on the BioSpe corpus

Although the evaluation result was good, the manual comparison of diverging sentences shows that a group of speculative sentences were not detected by BioExcom as it is shown in table 1 (the A2 ∩ spe sentences group).
The study of these sentences is considered as a starting point to improve the system performance. Speculative sentences of the A2 group are first analyzed then categorized according to their speculation linguistic markers. From this work, some new linguistic markers are added to existing rules or new rules are built.
For example, the indicator "*appear to/that*" of the sentence (12) is now recognized by BioExcom.

(12) "These different requirements for Dl and Ser appear to primarily result from their non- overlapping expression patterns rather than from distinct signaling properties".

The performance improvement of BioExcom system results are presented in table 5 and show that the system was able to automatically annotate 75,73% of the initially not detected sentences group after updating the linguistic resources but 24,27% of them were still not automatically detected due to the lack of the accurate linguistic markers (indicators or complementary clues).

[3] The BioSpe corpus contains (B) sentences group and speculative sentences from the (A1) and the (A2) sentences group.

[4] www.bio-excom.net/corpus.html

	Speculative sentences previously annotated by BioScope and not by BioExcom
Detected by BioExcom after improvement	287
Undetected by BioExcom after improvement	92
Total	379

Table 5: The BioExcom system performance improvement

Indeed, BioExcom was unable to annotate a group of speculative sentences dealing with the "*could*" indicator such as the sentence (13). In order to avoid noisy annotations, the BioExcom system only annotates a sentence with the *"could*" marker when there is conditionality or a possibility clue such as "*if*", "*whether*" since the "*could*" marker can express either the past form or the conditional form (in this case the sentence is speculative). In the sentence (13), there is no clue which could take off the ambiguity.

(13) "The observed patchy distribution could be caused by horizontal transfers and extinctions of Transib transposons in eukaryotic species".

Other sentences with the "*or*" indicator were also not annotated. Indeed, the presence of this marker in a sentence does not automatically mean that it deals with a speculation. For example, in sentence (14), the indicator "*or*" expresses a speculation whereas the sentence (15) (which also was annotated by BioScope) does not express a speculation.

(14) "By competition analysis with transcription factor consensus sequence oligonucleotides and by immunosupershift, transcription factor SP-1 or a closely related protein was shown to bind to this regulatory element".

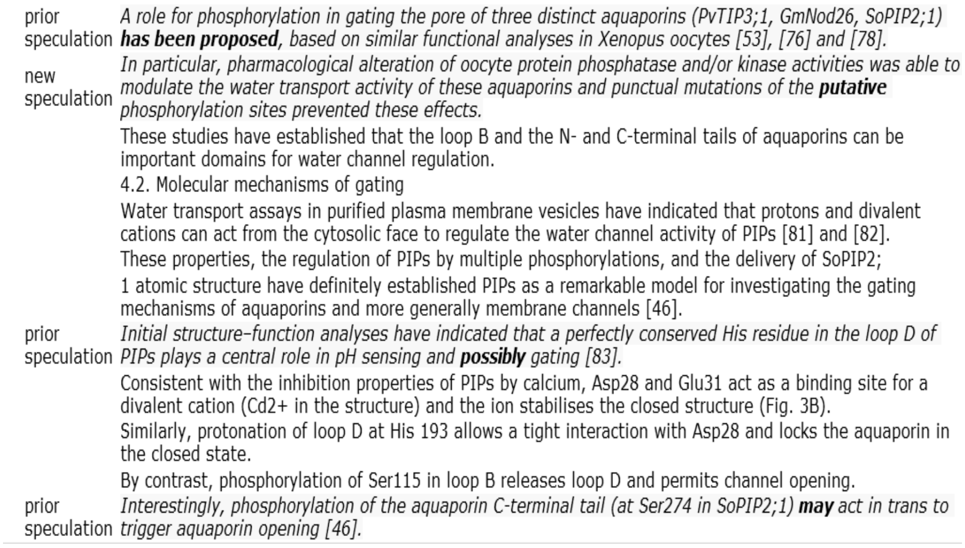
(15) "The CD34+ myelomonocytic cell line KG1 differentiates into dendritic-like cells in response to granmulocyte-macrophage CSF plus TNF-alpha, or PMA (with or without the calcium ionophore ionomycin, or TNF-alpha), with different stimuli mediating different aspects of the process".

5. An Online User Interface

We present a user interface[5] provided on line that is based on the improved (BioExcom_2) linguistic resources and aims to automatically annotate speculative sentences and to categorize them into "new" and "prior" speculation. The output text annotation is presented in figure 4.

6. Conclusion

The aim of our work was to evaluate the BioExcom system on a large scale. This task consisted to automatically annotating a part of the BioScope corpus. The comparison between the BioExcom automatic annotations and the BioScope manual annotations was useful to improve the BioExcom performance. A corpus BioSpe resulting from the evaluation task is built according the BioExcom speculation characterization. This corpus is made available on line and can be useful for machine learning systems. We have also presented in this study a user interface for the automatic annotation and categorization of speculative sentences.



prior speculation *A role for phosphorylation in gating the pore of three distinct aquaporins (PvTIP3;1, GmNod26, SoPIP2;1)* ***has been proposed****, based on similar functional analyses in Xenopus oocytes [53], [76] and [78].*

new speculation *In particular, pharmacological alteration of oocyte protein phosphatase and/or kinase activities was able to modulate the water transport activity of these aquaporins and punctual mutations of the* ***putative*** *phosphorylation sites prevented these effects.*

These studies have established that the loop B and the N- and C-terminal tails of aquaporins can be important domains for water channel regulation.

4.2. Molecular mechanisms of gating

Water transport assays in purified plasma membrane vesicles have indicated that protons and divalent cations can act from the cytosolic face to regulate the water channel activity of PIPs [81] and [82].

These properties, the regulation of PIPs by multiple phosphorylations, and the delivery of SoPIP2;

1 atomic structure have definitely established PIPs as a remarkable model for investigating the gating mechanisms of aquaporins and more generally membrane channels [46].

prior speculation *Initial structure–function analyses have indicated that a perfectly conserved His residue in the loop D of PIPs plays a central role in pH sensing and* ***possibly*** *gating [83].*

Consistent with the inhibition properties of PIPs by calcium, Asp28 and Glu31 act as a binding site for a divalent cation (Cd2+ in the structure) and the ion stabilises the closed structure (Fig. 3B).

Similarly, protonation of loop D at His 193 allows a tight interaction with Asp28 and locks the aquaporin in the closed state.

By contrast, phosphorylation of Ser115 in loop B releases loop D and permits channel opening.

prior speculation *Interestingly, phosphorylation of the aquaporin C-terminal tail (at Ser274 in SoPIP2;1)* ***may*** *act in trans to trigger aquaporin opening [46].*

Figure 4: Annotated text visualization (application output)

[5] http://www.bio-excom.net/acces.php

7. References

Alrahabi, M., Desclés, J.P.(2009). EXCOM: Plate-forme d'annotation sémantique de textes multilingues. In *Proceedings of the Natural Language Processing conference* Senlis, France, June 24-26.

Agarwal , S., Yu, H.(2010). Detecting Hedge Cues and their Scope in Biomedical Literature with Conditional Random Fields. *Journal of Biomedical Informatics 2010*, 43(6),pp. 953--961.

Ciccarese, P., Wu, E., Wong,G. Ocana, M. Kinoshita, J., Ruttenberg, A. Clark, T. (2008). The SWAN biomedical discourse ontology, *JBiomed* Inf. 41(5):739-51.

de Waard, A., Buckingham Shum, S., Carusi, A., Park, J., Samwald, M., and Sándor, Á. (2009). "Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims," In *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009),* co-located with the 8th Int'l Semantic Web Conference (ISWC-2009).

Desclés, J., Alrahabi, M., Desclés JP. (2009). BioExcom: Detection and Categorization of Speculative Sentences in Biomedical Literature, Human Language Technology. Challenges for Computer Science and Linguistics, 2009, *Lecture Notes in Computer Science*, Volume 6562/2011, pp 478--489.

Desclés, JP. (2006) Contextual Exploration Processing for Discourse Automatic Annotations of Texts. *In FLAIRS 2006*, invited speaker, Melbourne, Florida, pp 281-284

Djioua, B., Flores, JG., Blais, A., Desclés, JP., Guibert, G., Jackiewicz, A., Le Priol, F., Nait-Baha, L., Sauzay, B. (2006). EXCOM: an automatic annotation engine for semantic information. In *Proceeding of the nineteenth Florida Artificial Intelligence Research Society Conference FLAIRS 2006*, Melbourne, Florida, pp285--290

Friedman, C., Alderson, P., Austin, J., Cimino, J.J., Johnson, S.B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2): pp.161--174.

Hyland. K., (1998). *Hedging in Scientific Research Articles*. John Benjamins B.V

Kilicoglu, H., Bergler, S. (2008). Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics* 9 Suppl 11: S10

Light, M., Qiu, XY., Srinivasan, P. (2004). The Language of Bioscience: Facts, Speculations, and Statements in Between. In HLT-NAACL, ed, In *Proceedings in Workshop on Linking Biological Literature Ontologies And Database,* pp. 17--24

Medlock, B., Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of ACL* 2007.

Morante, R., Daelemans , W.(2009) Learning the scope of hedge cues in biomedical texts. In *Proceedings of the workshop on BioNLP. Boulder, Colorado: Association for Computational Linguistics;* 2009. pp. 28--36.

Szarvas, G.., Vincze, V., Farkas, R., Csirik, J. (2008). The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *In BioNLP ACL-2008* workshop.

Szarvas G,. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. In *Proceedings of ACL-08: HLT,* pp. 281--289, Colombus, Ohio, USA.

Thompson, P., Venturi, G., McNaught, J., Montemagni, S., Ananiadou, S. (2008). Categorising modality in biomedical texts. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical.Text Mining*. Marrakech, Morocco

Weakly Labeled Corpora as Silver Standard for Drug-Drug and Protein-Protein Interaction

Philippe Thomas[1*], **Tamara Bobić**[2,3*], **Martin Hofmann-Apitius**[2,3], **Ulf Leser**[1], **Roman Klinger**[2]

[1]Institute for Computer Science
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany

[2]Fraunhofer Institute for Algorithms
and Scientific Computing (SCAI)
Schloss Birlinghoven
53754 Sankt Augustin
Germany

[3]Bonn-Aachen Center for
Information Technology (B-IT)
Dahlmannstraße 2
53113 Bonn
Germany

{tbobic,klinger,hofmann-apitius}@scai.fraunhofer.de
{thomas,leser}@informatik.hu-berlin.de

Abstract

Relation extraction is frequently and successfully addressed by machine learning methods. The downside of this approach is the need for annotated training data, typically generated in tedious manual, cost intensive work. Distantly supervised approaches make use of weakly annotated data, which can be derived automatically. Recent work in the biomedical domain has applied distant supervision for protein-protein interaction (PPI) with reasonable results, by employing the IntAct database. Training from distantly labeled corpora is more challenging than from manually curated ones, as such data is inherently noisy. With this paper, we make two corpora publicly available to the community to allow for comparison of different methods that deal with the noise in a uniform setting. The first corpus is addressing protein-protein interaction (PPI), based on named entity recognition and the use of IntAct and KUPS databases, the second is concerned with drug-drug interaction (DDI), making use of the database DrugBank. Both corpora are in addition labeled with 5 state-of-the-art classifiers trained on annotated data, to allow for development of filter methods. Furthermore, we present in short our approach and results for distant supervision on these corpora as a strong baseline for future research.

Keywords: Distant Supervision, Relation Extraction, Silver Standard

1. Introduction

Relation Extraction (RE) in the biomedical domain is a discipline that is under extensive examination in the past decade, with a goal to automatically extract interacting pairs of entities from free text. Currently, a lot of relation extraction systems rely on machine learning, namely classifying pairs of entities to be related or not (Airola et al., 2008; Miwa et al., 2009; Kim et al., 2010). Despite the fact that machine learning has been most successful in identifying relevant relations in text, a drawback is the need for manually annotated training data. Domain experts have to dedicate time and effort to this tedious and labor-intensive process.

As a consequence of the overall scarcity of annotated corpora for relation extraction in the biomedical domain, the approach of distant supervision, *e. g.* automatic labeling of a training set is emerging. Many approaches follow the distant supervision assumption (Mintz et al., 2009; Riedel et al., 2010): "If two entities participate in a relation, all sentences that mention these two entities express that relation." Obviously, this assumption does not hold in general, and therefore exceptions need to be detected.

To allow the community to compare different approaches for distant supervision, we make two corpora, one for protein-protein interaction (PPI) and one for drug-drug interaction (DDI) publicly available.[1] In addition, we present our results on this task as a strong baseline. To complete the purpose of a silver standard, annotations of well-established supervised models on this corpus are included.

1.1. Related Work

Distant supervision approaches have received considerable attention in the past few years. However, most of the work is focusing on domains other than biomedical texts. Mintz et al. (2009) use distant supervision to learn to extract relations that are represented in Freebase (Bollacker et al., 2008). Yao et al. (2010) use Freebase as a source of supervision, dealing with entity identification and relation extraction in a joint fashion. Riedel et al. (2010) argue that distant supervision leads to noisy training data that hurts precision and suggest a two step approach to reduce this problem. Vlachos et al. (2009) tackle the problem of biomedical event extraction. The scope of their interest is to identify different event types without using a knowledge base as a source of supervision, but explore the possibility of inferring relations from the text based on the trigger words and dependency parsing, without previously annotated data. Thomas et al. (2011b) make use of a distantly labeled corpus for protein-protein interaction extraction. Different strategies are evaluated to select informative training instances. Buyko et al. (2012) examine the usability of knowledge from a database to generate training sets that capture gene-drug, gene-disease and drug-disease relations.

The CALBC project asks for automated annotation of entity classes in a common corpus to generate a silver standard by combining different predictions (Rebholz-Schuhmann and Ş. Kafkas, 2011). The usability of automatically derived corpora has been recently demonstrated for the task of noun-phrase chunking (Kang et al., 2012). The EVEX data set is the result of applying named entity recognition, parsing and event extraction on full MEDLINE (Landeghem et al., 2011).

[*]These authors contributed equally.

[1]These two corpora are publicly at: http://www.scai.fraunhofer.de/ppi-ddi-silverstandard.html.

Corpus	Positive pairs	Negative pairs	Total
AIMed	1000 (0.17)	4,834 (0.82)	5,834
BioInfer	2,534 (0.26)	7,132 (0.73)	9,666
HPRD50	163 (0.38)	270 (0.62)	433
IEPA	335 (0.41)	482 (0.59)	817
LLL	164 (0.49)	166 (0.50)	330
DDI train	2,400 (0.10)	21,411 (0.90)	23,811
DDI test	755 (0.11)	6,275 (0.89)	7,030

Table 1: Basic statistics of the five PPI and two DDI corpora. Ratios are given in brackets.

1.2. Interaction Databases

The IntAct database (Kerrien et al., 2012) contains protein-protein interaction information. It consists of 290,947 binary interaction evidences, including 39,235 unique pairs of interacting proteins for human species.[2] KUPS (Chen et al., 2010) is a database that combines entries from three manually curated PPI databases (IntAct, MINT (Chatr-aryamontri et al., 2007) and HPRD50 (Prasad et al., 2009)) and contains 185,446 positive pairs from various model organisms, out of which 69,600 belong to human species.[3] Enriching IntAct interaction information with the KUPS database leads to 57,589 unique pairs.[4]

The database DrugBank (Knox et al., 2011) combines detailed drug data with comprehensive drug target information. It consists of 6,707 drug entries. Apart from information about its targets, for certain drugs known interactions with other drugs are given. Altogether, we obtain 11,335 unique DDI pairs.

1.3. Manually Curated Corpora

Pyysalo et al. (2008) made five corpora for protein-protein interaction available in the same XML-based file format. Their properties, like size and ratio of positive and negative examples, differ greatly, the latter being the main cause of performance differences when evaluating on these corpora. Moreover, annotation guidelines and contexts differ: AIMed (Bunescu et al., 2005) and HPRD50 (Fundel et al., 2007) are human-focused, LLL (Nedellec, 2005) on Bacillus subtilis, BioInfer (Pyysalo et al., 2007) contains information from various organisms, and IEPA (Ding et al., 2002) is made of sentences that describe 10 selected chemicals, majority of which are proteins, and their interactions.

Segura-Bedmar et al. (2011b) published a drug-drug interaction corpus where the drug mentions have been automatically detected with MetaMap and their pair-wise relations are manually annotated. The corpus is divided into a training and testing set, generated from web-documents describing drug effects.

An overview of the corpora is given in Table 1.

2. Methods

In this section, the workflow to prepare the two corpora is presented.

2.1. Automatically Labeling a Corpus

One of the most important source of publications in the biomedical domain is MEDLINE[5], currently containing more than 21 million citations.[6] The initial step is annotation of named entities and entity normalization against the databases mentioned in Section 1.2. – in our case performed by ProMiner (Hanisch et al., 2005), a tool proving state-of-the-art results in *e. g.* the BioCreative competition (Fluck et al., 2007). Based on the named entity recognition, only sentences containing co-occurrences of relevant entities are further processed. Based on the distant supervision assumption, each pair of entities is labeled as related if mentioned so in a structured interaction database. Following the closed world assumption, all remaining entity pairs are labeled as non-interacting. To avoid information leakage and biased classification, all documents which are contained in the test corpus are removed from the distantly labeled corpus. Each corpus is sub-sampled to a size of 200,000 entity-pairs, which is more than an order of magnitude larger than any manually annotated PPI or DDI corpus.

2.2. Corpus Preprocessing

Sentences are parsed using the Charniak-Lease parser (Lease and Charniak, 2005) with a self-trained re-ranking model specialized for biomedical texts (McClosky, 2010). Resulting constituent parse trees are converted into dependency graphs using the Stanford converter (Marneffe et al., 2006). We create an augmented XML following the recommendations of Airola et al. (2008). This XML encompasses tokens with respective part-of-speech tags, constituent parse tree, and dependency parse tree information. The pairs are augmented with class labels predicted from five different relation extraction methods (see Section 2.3.). For interacting pairs in the PPI corpus we provide the original source (IntAct or KUPS) along with the information if the pair is made of self-interacting proteins. For sentences of the PPI corpus we include the information if an interaction (trigger) word is present. However, in case of DDI trigger-based filtering is not applied (see Bobić et al. (2012)).

2.3. Pair Annotation

Labeling two large corpora with database knowledge is the main contribution of this paper. Additionally, we supplement the corpus with predictions of five state-of-the-art relation extraction approaches to provide a supplementing layer of information. (An assessment of the used methodologies for relation extraction was performed by Tikk et al. (2010).)

This includes the shallow linguistic (SL) (Giuliano et al., 2006), all-paths graph (APG) (Airola et al., 2008), subtree (ST) (Vishwanathan and Smola, 2002), subset tree SST (Collins and Duffy, 2001), and spectrum tree (SpT) (Kuboyama et al., 2007) method, which exploit different views on the data. Parameter optimization was performed as

[2] As of January 27th, 2012.

[3] As of August 16th, 2010.

[4] 45,684 out of 69,600 human PPI pairs are available from the KUPS web service due to computational and storage limitations (personal communication).

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

[6] As of January, 2012.

described by Tikk et al. (2010). For a detailed description of the feature setting and approach, we refer to the original publications. Entities were blinded by replacing the entity name with a generic string to ensure the generality of the approach. Constituent parse trees have been reduced to the shortest-enclosed parse following the recommendations from Zhang et al. (2006). All five methods are trained on the union of all five PPI corpora and the DDI training and test set respectively. Note that the predictions coming from the five methods are biased towards these training corpora: Models trained on the resulting silver standard (excluding the database annotation) are likely to obtain a too optimistic result, even though the respective sentences from the test set are not used in the training process.

3. Results

In this section, we start with an overview of state-of-the-art results for fully supervised relation extraction on PPI and DDI corpora (see Table 1). Section 3.2. gives a statistical outline of the two distantly labeled corpora. Subsequently we present the results of the five relation extraction methods trained on manually annotated data and applied on the distantly labeled corpora. Finally, we present our results for models trained on distantly labeled PPI and DDI data, when evaluated on manually annotated corpora, as a strong baseline for future research.

3.1. Performance Overview of Supervised RE Systems

Protein-protein interactions have been extensively investigated in the past decade because of their biological significance. Machine learning approaches have shown the best performance in this domain (*e. g.* BioNLP (Cohen et al., 2011; Tsujii et al., 2011) and DDIExtraction Shared Task (Segura-Bedmar et al., 2011a)).

Our relation extraction system is based on the linear support vector machine classifier LibLINEAR (Fan et al., 2008). The approach employs lexical and dependency parsing features, as explained by Bobić et al. (2012).

Table 4 shows a comparison of state-of-the-art relation extraction systems' performances on 5 PPI corpora, determined by document level 10-fold cross-validation. In Table 2, results of the five best performing systems on the DDI test data set of the DDI extraction workshop are shown. Note that the first three systems use ensemble based methods combining the output of several different classifiers. In addition, the performance of our system, which is later used for distant supervision, is shown in both tables.

3.2. Distantly Labeled Corpora for DDI and PPI

The file format of the corpora is by large self explanatory and strongly follows an established file format (Airola et al., 2008; Pyysalo et al., 2008). A short excerpt of the DDI corpus is shown in the appendix. The example consists of one sentence with two annotated drugs that participate in a relation according to DrugBank.

Basic statistics of the two distantly labeled corpora are shown in Table 3. The Charniak-Lease parser does not produce results for nine sentences in the PPI corpus and 14 sentences in the DDI corpus. In general, most methods fail to predict class labels for instances contained in these sentences, leading to a reduced number of predictions per corpus. However, the effect is only marginal as <1 % of all entity pairs are affected by this problem.

Methods	P	R	F_1
Thomas et al. (2011a)	60.5	**71.9**	**65.7**
Chowdhury et al. (2011)	58.6	70.5	64.0
Chowdhury and Lavelli (2011)	58.4	70.1	63.7
Björne et al. (2011)	58.0	68.9	63.0
Minard et al. (2011)	55.2	64.9	59.6
Our system (*lex*)	62.7	52.1	56.9
Our system (*lex+dep*)	**66.9**	57.9	62.1

Table 2: Comparison of fully supervised relations extraction systems for DDI. (*lex* denotes the use of lexical features, *lex+dep* the additional use of dependency parsing-based features.) The first three systems are based on ensemble learning.

	PPI	DDI
Abstracts	49,958	76,859
Sentences	51,934	79,701
Pos. Sent.	19,891	5,587
Tokens	1,608,899	2,520,545
Entities	150,886	203,315
Pairs	200,000	200,000
Pos. Pairs	37,600	8,705

Table 3: Statistics of the distant PPI and DDI corpora. (pos. sent. denotes the number of sentences with at least one related entity pair.)

3.3. Pair Annotation

As shown in Table 5, relation extraction methods tend to classify between 10.9 % and 16.8 % of all protein pairs as interacting. However, the overall ratio of positive instances across all five PPI corpora is greater, measuring up to 32.6 %. We observe similar values for the distant DDI corpus with ratios ranging from 12.7 % to 19.6 %.

The distribution of confidence scores (distance to the hyperplane) for all methods on both corpora is shown in Figure 1. Instances with a negative sign are classified as non-interacting and instances with a positive sign are classified as interacting. The linear association between different methods is assessed using Pearson correlation for all instances contained in the distantly supervised corpus. We observe correlation coefficients ranging from 0.29 (APG versus SpT) to 0.59 (APG versus SL) for PPI and between 0.34 (APG vs ST) to 0.71 (ST vs SST) for DDI. Significance of all pairwise correlations is assessed using a t-test and is in all cases highly significant (p-value < 0.01). Correlation is exemplarily depicted as scatterplot for SL and APG on PPI in Figure 2. Both methods agree on the predicted class label on instances contained in the first and third quadrant, whereas the two methods have conflicting results for instances in the second and fourth quadrant. The figure indicates that some instances can be confidently classified by one method

	AIMed			BioInfer			HPRD50			IEPA			LLL		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Airola et al. (2008)	52.9	61.8	56.4	56.7	67.2	61.3	64.3	65.8	63.4	69.6	**82.7**	**75.1**	72.5	87.2	76.8
Kim et al. (2010)	61.4	53.2	56.6	61.8	54.2	57.6	66.7	69.2	67.8	**73.7**	71.8	72.9	76.9	91.1	**82.4**
Fayruzov et al. (2009)			39.0			34.0			56.0			72.0			76.0
Liu et al. (2010)			54.7			59.8			64.9			62.1			78.1
Miwa et al. (2009)	55.0	**68.8**	**60.8**	65.7	**71.1**	**68.1**	68.5	**76.1**	70.9	67.5	78.6	71.7	**77.6**	86.0	80.1
Tikk et al. (2010)	47.5	65.5	54.5	55.1	66.5	60.0	64.4	67	64.2	71.2	69.3	69.3	74.5	85.3	74.5
Our system (*lex*)	62.9	50.0	55.7	59.3	55.1	57.1	**72.4**	75.6	**73.9**	67.7	73.3	70.4	66.6	88.6	76.1
Our system (*lex+dep*)	**63.6**	52.0	57.2	**65.8**	62.9	64.3	70.8	74.0	72.4	70.4	76.1	73.2	70.4	**91.6**	79.6

Table 4: Comparison of fully supervised relation extraction systems for PPI.

	PPI		DDI	
Method	positive	negative	positive	negative
SL	33,677 (16.8)	166,219	25,344 (12.7)	174,539
SpT	21,971 (10.9)	177,921	29,324 (14.6)	170,558
ST	28,885 (14.4)	171,112	39,286 (19.6)	160,597
SST	24,840 (12.4)	175,157	25,841 (12.9)	174,039
APG	26,313 (13.1)	173,686	25,357 (12.7)	174,643

Table 5: Distribution of positive and negative instances for the different methods on both distantly labeled corpora. The ratio of positive examples is given in brackets.

(high distance to the hyperplane), but the other method is comparably inconfident. This suggests a great variability between the methods.

Even though the correlation between the methods is lower than expected, the inter-classification agreement (accuracy) is comparably high and ranges between 80.7 % to 86.4 % and 78.2 % to 84.6 % for all PPI and DDI instances respectively. We observe a large agreement between the distantly labeled corpus and the classification methods with approximately 76 % overall agreement for PPI and 80 % for DDI. The association between distantly labeled corpora and all classification methods is significant according to a fisher test (p-value < 0.01), except for SpT where we observe a p-value of 0.04. However, the large overall agreement is due to the high number of negative instances in the distant corpora and predicted by the different methods. For positive PPI instances alone we observe an agreement of approximately 27 % between instances labeled as interacting by our knowledge base and the classification methods. Similar effects can be observed for the DDI corpus. We assessed the overall agreement between methods and the two distantly labeled corpora using Cohens κ. For PPI we observe values ranging between 0.07 to 0.19 and for DDI we observe κ values of 0.03. The low κ values show a comparably small agreement between classification methods and distantly labeled corpora and more sophisticated filtering techniques might be required to make optimal use of the corpus. Results in terms of precision, recall and F_1 can be seen in Table 6.

	PPI			DDI		
Method	P	R	F_1	P	R	F_1
SL	35.1	31.4	33.2	6.4	18.7	9.5
SpT	27.4	16.0	20.2	4.5	15.3	7.0
ST	35.2	27.1	30.6	5.5	25.1	9.1
SST	32.3	21.4	25.7	6.2	18.6	9.3
APG	36.0	25.1	29.6	5.8	16.7	8.6

Table 6: Comparison of all methods on both distantly labeled corpora. (P denotes precision, R recall and F_1 the harmonic mean of P and R)

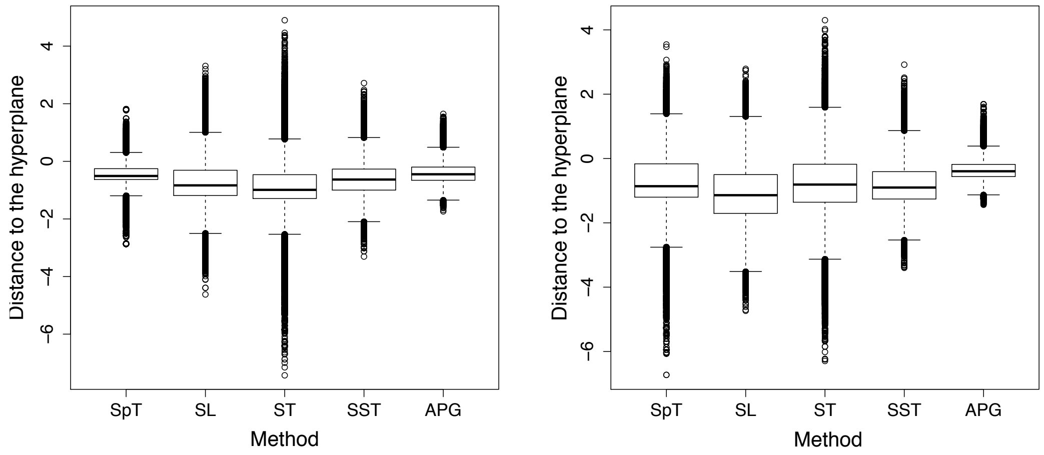
3.4. Baselines for Distantly Supervised Models

For each experiment we sample random subsets of 10,000 entity pairs from the proposed corpora. All experiments are performed five times to reduce the influence of sampling different subsets. We apply the method proposed by Bobić et al. (2012), with dependency parsing based features and filtering auto-interacting entities. For PPI, trigger-based filtering is applied (compare to Section 2.2.). Table 7 shows the average performance trained on the distantly labeled PPI and DDI corpora.

Note that the instance labels used for training the model are based solely on database knowledge. The information provided by five supervised methods (addressed in Section 2.3.) are not taken into account for generating baseline results, although they are available to be used in future work.

Our system outperforms co-occurrence results for all five PPI corpora, as shown in Table 7. F_1 measure of AIMed and BioInfer, for which we assume to have the most realistic pos/neg ratio, outperforms the baseline by around 9 percentage points (pp). HPRD50, IEPA and LLL have an improvement of 4.7 pp, 5.3 pp and 0.8 pp respectively, due to high fractions of positive instances (leading to a strong co-occurrence baseline).

Evaluation on corpora that have different properties than the training set leads to decreased performance (Airola et al., 2008; Tikk et al., 2010). Often, the properties of a test corpus (like MEDLINE) are not known for real world



(a) PPI

(b) DDI

Figure 1: Boxplot on distance to the hyperplane of all used methods for both corpora.

	Our system (*lex*)			Our system (*lex+dep*)			Co-occ.			Tikk et al. (2010)		
Corpus	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
AIMed	25.6	78.4	38.6	25.0	81.9	38.4	17.1	100	29.3	28.3	86.6	**42.6**
BioInfer	40.4	66.7	**50.3**	40.3	66.9	**50.3**	26.2	100	41.5	62.8	36.5	46.2
HPRD50	45.7	85.1	59.4	44.9	86.3	59.0	37.6	100	54.7	56.9	68.7	**62.2**
IEPA	50.0	87.2	**63.5**	49.9	85.8	63.1	41.0	100	58.2	71.0	52.5	60.4
LLL	56.4	83.1	**67.2**	56.3	83.2	**67.2**	49.7	100	66.4	79.0	57.3	66.4
DDI	33.2	39.2	36.0	33.0	44.1	**37.7**	10.7	100	19.4	—	—	—

Table 7: Results (in %) achieved when training on 10,000 distantly labeled instances and testing on 5 PPI corpora and the DDI test corpus, respectively.

applications. Thus cross-learning[7] is considered to provide a more realistic scenario to compare the performance of distantly supervised systems to fully supervised systems. Our approach outperforms the state-of-the-art cross-learning results from Tikk et al. (2010) in three out of five corpora, most notably in case of BioInfer where an increase of more than 4 pp in F_1 measure is observable.

In the case of drug-drug interaction, it is noteworthy that the manually annotated corpora are generated from web documents discussing drug effects which are not necessarily contained in MEDLINE. Hence, this evaluation corpus can be considered as out-domain and provides additional insights on the robustness of distant-supervision. Table 7 shows that compared to co-occurrence, a gain of more than 18 pp is achieved when training on a distantly labeled DDI corpus. Taking into account the high class imbalance of the DDI test set (see Table 1), which is most similar to the AIMed corpus, a F_1 measure of 37.7 % is encouraging.

Application of distant supervision to five substantially different PPI corpora and further utilization of the same workflow to DDI confirms its robustness and usability.

4. Discussion

This paper introduces two distantly labeled corpora created for the purpose of protein-protein and drug-drug interaction extraction. Corpus generation and the process of automatic pair labeling using database information are presented, together with strong baseline results for distantly supervised relation extraction.

In addition to entity-pair annotation based on a knowledge base, we add predictions from five relation extraction systems, trained on manually annotated corpora. These annotations can be exploited to develop better instance filtering techniques. Several assessments demonstrated the superiority of ensemble methods, hence it might be beneficial to combine classifier predictions for the sake of higher method robustness.

Our distant supervision baseline achieves competitive results and outperforms co-occurrence in all test cases. Comparison to fully supervised cross-learning results for PPI argues for the opportunities of using automatically annotated data.

This paper presents the potential of distant learning to allow a fully automated relation extraction process. The PPI and DDI corpora are made freely available to the community such that novel strategies of efficient employment of database knowledge can be compared.

[7] For five PPI corpora: train on four, test on the remaining.

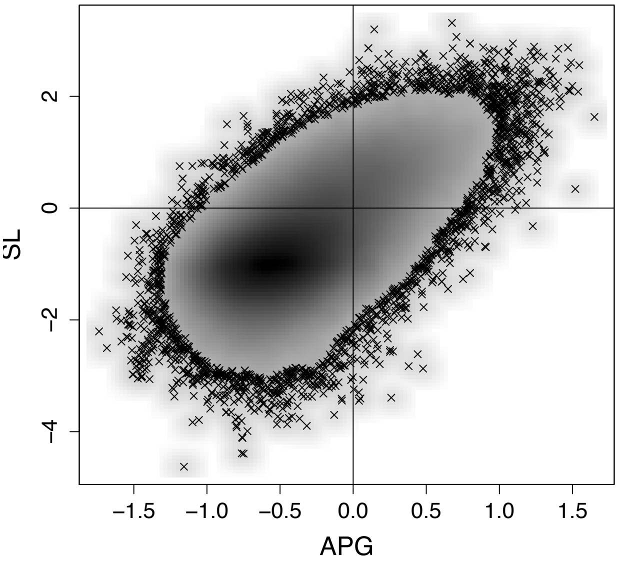


Figure 2: Scatterplot for distance to the hyperplane between APG and SL on the distantly labeled PPI corpus. Warm regions (dark) indicate an accumulation of instances whereas light regions contain no instances. The 2,000 points in areas with lowest regional density are plotted separately.

5. Acknowledgements

Thanks to our colleagues Juliane Fluck and Theo Mevissen for fruitful discussions and technical help. T. Bobić was partially funded by the Bonn-Aachen International Center for Information Technology (B-IT) Research School. P. Thomas was funded by the German Federal Ministry of Education and Research (grant No 0315417B). R. Klinger was partially funded by the European Community's Seventh Framework Programme [FP7/2007-2011] under grant agreement no. 248726.

6. References

A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. 2008. All-paths Graph Kernel for Protein-protein Interaction Extraction with Evaluation of Cross-corpus Learning. *BMC Bioinformatics*, 9(Suppl 11):S2.

J. Björne, A. Airola, T. Pahikkala, and T. Salakoski. 2011. Drug-drug interaction extraction with RLS and SVM classiffers. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 35–42.

T. Bobić, R. Klinger, P. Thomas, and M. Hofmann-Apitius. 2012. Improving Distantly Supervised Extraction of Drug-Drug and Protein-Protein Interactionsp. In O. Abend, C. Biemann, A. Korhonen, A. Rappoport, R. Reichart, and A. Sgaard, editors, *ROBUS-UNSUP*.

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pages 1247–1250.

R. C. Bunescu, R. Ge, R. J. Kate, E. M. Marcotte, R. J. Mooney, A. K. Ramani, and Y. Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med*, 33(2):139–155, Feb.

E. Buyko, E. Beisswanger, and U. Hahn. 2012. The extraction of pharmacogenetic and pharmacogenomic relations–a case study using pharmgkb. *PSB*, pages 376–387.

A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M.V. Schneider, L. Castagnoli, and G. Cesareni. 2007. MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 35(Database issue):D572–D574.

X. Chen, J. C. Jeong, and P. Dermyer. 2010. KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions. *Nucleic Acids Res*, 39(Database issue):D750–D754.

F. M. Chowdhury and A. Lavelli. 2011. Drug-drug interaction extraction using composite kernels. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 27–33.

F. M. Chowdhury, A. B. Abacha, A. Lavelli, and P. Zweigenbaum. 2011. Two different machine learning techniques for drug-drug interaction extraction. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 19–26.

K. B. Cohen, D. Demner-Fushman, S. Ananiadou, J. Pestian, J. Tsujii, and B. Webber, editors. 2011. *Proceedings of the BioNLP*.

M. Collins and N. Duffy. 2001. Convolution Kernels for Natural Language. In *Proc. of Neural Information Processing Systems (NIPS'01)*, pages 625–632, Vancouver, BC, Canada.

J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. 2002. Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, pages 326–337.

E. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Machine Learning Research*, 9:1871–1874.

T. Fayruzov, M. De Cock, C. Cornelis, and V. Hoste. 2009. Linguistic feature analysis for protein interaction extraction. *BMC Bioinformatics*, 10(1):374.

J. Fluck, H. T. Mevissen, H. Dach, M. Oster, and M. Hofmann-Apitius. 2007. ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. In *BioCreative 2*, pages 149–151.

K. Fundel, R. Kuffner, and R. Zimmer. 2007. RelEx-Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

C. Giuliano, A. Lavelli, and L. Romano. 2006. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. In *Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 401–408, Trento, Italy.

D. Hanisch, K. Fundel, H. T. Mevissen, R. Zimmer, and J. Fluck. 2005. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S14.

N. Kang, E. M. van Mulligen, and J. A. Kors. 2012. Training text chunkers on a silver standard corpus: can silver replace gold? *BMC Bioinformatics*, 13(1):17, Jan.

S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, C. Jandrasits, R.C. Jimenez, J. Khadake, U. Mahadevan, P. Masson, I. Pedruzzi, E. Pfeiffenberger,

P. Porras, A. Raghunath, B. Roechert, S. Orchard, and H. Hermjakob. 2012. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*, 40:D841–D846.

S. Kim, J. Yoon, J. Yang, and S. Park. 2010. Walk-weighted subsequence kernels for protein-protein interaction extraction. *BMC Bioinformatics*, 11:107.

C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. Chi Guo, and D.S Wishart. 2011. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*, 39(Database issue):D1035–D1041.

T. Kuboyama, K. Hirata, H. Kashima, K. F. Aoki-Kinoshita, and H. Yasuda. 2007. A Spectrum Tree Kernel. *Information and Media Technologies*, 2(1):292–299.

S. Van Landeghem, F. Ginter, Y. Van de Peer, and T. Salakoski. 2011. EVEX: A PubMed-Scale Resource for Homology-Based Generalization of Text Mining Predictions. In *Proc. of BioNLP*, pages 28–37.

M. Lease and E. Charniak. 2005. Parsing biomedical literature. In *Proc. of IJCNLP'05*, pages 58–69.

B. Liu, L. Qian, H. Wang, and G. Zhou. 2010. Dependency-driven feature-based learning for extracting protein-protein interactions from biomedical text. In *COLING*, pages 757–765.

M. C. De Marneffe, B. Maccartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*, pages 449–454.

D. McClosky. 2010. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Brown University.

A. L. Minard, L. Makour, A. L. Ligozat, and B. Grau. 2011. Feature Selection for Drug-Drug Interaction Detection Using Machine-Learning Based Approaches. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 43–50.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP*, pages 1003–1011.

M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii. 2009. A Rich Feature Vector for Protein-Protein Interaction Extraction from Multiple Corpora. In *Proc. of EMNLP*, pages 121–130.

C. Nedellec. 2005. Learning language in logic-genic interaction extraction challenge. In *Proc. of the ICML05 workshop: Learning Language in Logic (LLL'05)*, volume 18, pages 97–99.

T. S. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A.Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey. 2009. Human Protein Reference Database–2009 update. *Nucleic Acids Res*, 37(Database issue):D767–D772.

S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. 2007. BioInfer: A Corpus for Information Extraction in the Biomedical Domain. *BMC Bioinformatics*, 8(50).

S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein–protein interaction corpora. *BMC Bioinformatics*, 9 Suppl 3:S6.

D. Rebholz-Schuhmann and Ş. Kafkas, editors. 2011. *Proceedings of the Second CALBC Workshop*.

S. Riedel, L. Yao, and A. McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *ECML PKDD*.

I. Segura-Bedmar, P. Martínez, and D. Sanchez-Cisneros, editors. 2011a. *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*.

I. Segura-Bedmar, P. Martínez, and D. Sanchez-Cisneros. 2011b. The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. In *Challenge Task on Drug-Drug Interaction Extraction 2011*, pages 1–9.

P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser. 2011a. Relation Extraction for Drug-Drug Interactions using Ensemble Learning. In *Challenge Task on Drug-Drug Interaction Extraction*, pages 11–18.

P. Thomas, I. Solt, R. Klinger, and U. Leser. 2011b. Learning Protein Protein Interaction Extraction using Distant Supervision. In *Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pages 34–41.

D. Tikk, P. Thomas, P. Palaga, J. Hakenberg, and U. Leser. 2010. A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature. *PLoS Comput Biol*, 6:e1000837.

J. Tsujii, J.-D. Kim, and S. Pyysalo, editors. 2011. *Proceedings of the BioNLP Shared Task*.

S. V. N. Vishwanathan and A. J. Smola. 2002. Fast Kernels for String and Tree Matching. In *Proc. of Neural Information Processing Systems (NIPS'02)*, pages 569–576, Vancouver, BC, Canada.

A. Vlachos, P. Buttery, D. Ó Séaghdha, and T. Briscoe. 2009. Biomedical Event Extraction without Training Data. In *BioNLP*, pages 37–40.

L. Yao, S. Riedel, and A. McCallum. 2010. Collective Cross-Document Relation Extraction Without Labeled Data. In *EMNLP*, pages 1013–1023.

M. Zhang, J. Zhang, J. Su, and G. Zhou. 2006. A Composite Kernel to Extract Relations between Entities with Both Flat and Structured Features. In *Proc. of the 21st International Conference on Computational Linguistics*, pages 825–832, Sydney, Australia, July.

7. Appendix

An excerpt of the corpus in XML format:

```
<corpus source=”SilverDDICorpus”>
  <document id=”d3” origId=”10796253”>
    <sentence id=”d3.s0” origId=”10796253.s14”
      text=”In the subset with initial BUN/creatinine ratio > 20 mg/mg, 2 of 18 patients receiving furosemide
       could not complete a 3−dose course of indomethacin because of toxicity.”>
      <entity charOffset=”87−96” id=”d3.s0.e0” origId=”10796253.s14.e0” text=”furosemide”
      type=”drug”/>
      <entity charOffset=”136−147” id=”d3.s0.e1” origId=”10796253.s14.e1” text=”indomethacin”
      type=”drug”/>
      <pair e1=”d3.s0.e0” e2=”d3.s0.e1” id=”d3.s0.p0” interaction=”True” source=”DrugBank”
      APG=”0.32” SL=”0.60” ST=”−1.08” SST=”0.12” SpT=”0.34”/>
      <sentenceanalyses>
        <tokenizations>
          <tokenization tokenizer=”Charniak−Lease”>
          <token POS=”IN” charOffset=”0−1” id=”t_1” text=”In”/>
          <token POS=”DT” charOffset=”3−5” id=”t_2” text=”the”/>
          <token POS=”NN” charOffset=”7−12” id=”t_3” text=”subset”/>
          ...
          </tokenization>
        </tokenizations>
      <bracketings>
        <bracketing tokenizer=”Charniak−Lease” parser=”Charniak−Lease” bracketing=”(S1 (S (S (PP
        (IN In) (NP (NP (DT the) (NN subset)) (PP (IN with) (NP (NP (JJ initial) (NN BUN/creatinine) (NN ratio) (NN &gt;))
        (NP (CD 20) (NN mg/mg)))))) (, ,) (NP (NP (CD 2)) (PP (IN of) (NP (NP (CD 18) (NNS patients)) (VP (VBG receiving)
        (NP (NN furosemide)))))) (VP (MD could) (RB not) (VP (VB complete) (NP (NP (DT a) (JJ 3−dose) (NN course))
        (PP (IN of) (NP (NN indomethacin)))) (PP (IN because) (IN of) (NP (NN toxicity)))))) (. .)))”>
         <charOffsetMapEntry sentenceTextCharOffset=”0−1” bracketingCharOffset=”18−19”/>
         <charOffsetMapEntry sentenceTextCharOffset=”3−5” bracketingCharOffset=”34−36”/>
         <charOffsetMapEntry sentenceTextCharOffset=”7−12” bracketingCharOffset=”43−48”/>
         ...
        </bracketing>
      </bracketings>
      <parses>
        <parse tokenizer=”Charniak−Lease” parser=”Charniak−Lease”>
          <dependency id=”d_1” t1=”t_3” t2=”t_2” type=”det” origId=”det(subset−3, the−2)”/>
          <dependency id=”d_2” t1=”t_20” t2=”t_3” type=”prep_in” origId=”prep_in(complete−20, subset−3)”/>
          <dependency id=”d_3” t1=”t_8” t2=”t_5” type=”amod” origId=”amod(&gt;−8, initial−5)”/>
          ...
        </parse>
      </parses>
     </sentenceanalyses>
    </sentence>
      ...
  </document>
  ...
</corpus>
```

Developing Specifications for Light Annotation Tasks in the Biomedical Domain

Amber Stubbs

Laboratory for Linguistics and Computation
Department of Computer Science
Brandeis University, Waltham, MA 02453
astubbs@cs.brandeis.edu

Abstract

Biomedical texts pose an interesting challenge in natural language processing tasks. While the information contained in them is important to people of all backgrounds, often they are stylistically complex with specialized vocabularies, and require advanced degrees or other special training to interpret correctly. Because of this, researchers in Natural Language Processing are often at a disadvantage when it comes to extracting task-specific information from these texts: the experts who are best able to understand them may not have the time or interest in completing complicated and time-consuming annotations for use in corpus analysis and machine learning. This paper proposes a methodology for creating light annotation tasks for biomedical corpora that can be used to create useful annotations without requiring extensive training or exceptionally long annotation periods. The utility of the proposed methodology is examined in light of existing annotation projects, as well as through the lens of a case study using hospital discharge summaries for patient selection based on eligibility criteria.

Keywords: annotation, biomedical, methodology

1. Introduction

Text mining of biomedical corpora is a field that has been growing rapidly over the past decade. However, complex biomedical texts offer a unique challenge to computational linguists, who may not always have the domain-specific knowledge required to fully understand and interpret the texts from which they wish to mine information. At the same time, the people who do have this knowledge (doctors, biologists, chemists, etc.) may not have the time or inclination to provide sufficient professional information and linguistic insight to help researchers create useful datasets for training machine learning algorithms. Additionally, hiring such experts as consultants in order for them to perform annotations can be prohibitively expensive.

Naturally, not every query into biomedical texts requires the help of an MD or biologist— part-of-speech tagging, for instance, can generally be done by native speakers of the language even when the vocabulary is unfamiliar. Similarly, named entity and event annotations also do not always require domain knowledge, unless a terminologically rich annotation system is being used. However, as the field of biomedical text mining and information extraction expands, the questions being asked about the data begin to move from "Which of these words are nouns?" and "Which of these are events?" to "Which of these indicate disease X?" and "What are the temporal relationships between these events?" These questions are less easily answered by computational linguists, and more often require domain-specific knowledge and/or training to be properly addressed.

Light annotation tasks[1] are, in theory, an ideal way of solving this dilemma of linguistic complexity versus expert understanding of the literature, as they can exploit information about the chosen corpus without requiring full linguistic annotation. However, it is not easy to create an annotation task that is light (in terms of work required to obtain the annotation, both physically and mentally) and contentful (in terms of later utility). Ideally, a light annotation is acquired for a particular question or corpus, and then additional relevant information (part-of-speech tagging, semantic roles, document structure) is added later as a way of providing more features to the machine learning algorithms.

While light annotation tasks themselves are not new in the biomedical domain—some parts of past i2b2 (Informatics for Integrating Biology and the Bedside) challenges (Uzuner et al., 2007) have relied on them for creating datasets that are augmented by challenge participants, BioNLP tasks have benefited from simplifying annotation tasks used for other purposes (Kim et al., 2009; Kim et al., 2011), and systems like the Automated Retrieval Consol (ARC) (D'Avolio et al., 2010; D'Avolio et al., 2011) use them for data mining, for instance—no methodology or desiderata has been proposed to date for creating meaningful light annotation tasks.

This paper introduces such a methodology, which can be used in conjunction with current standards in corpus and computational linguistics. It is meant to be used in relation to the MATTER (Model, Annotate, Train, Test, Evaluate, Revise) cycle, a development cycle for annotation tasks (Pustejovsky, 2006; Pustejovsky and Stubbs, 2012). At the end of the paper, a case study using this methodology is examined, which focuses on using expert knowledge to create an annotation that represents patients who meet selection criteria for a medical study based on their hospital discharge summaries.

[1] For the purposes of this paper, a 'light annotation' is a textual markup that uses tags that are under-specified in terms of linguistic content, generally for the purpose of creating a task that requires less work to complete. This is in contrast to shallow annotation, such as when a shallow syntactic parse is performed over sentence structures.

2. Related Work

There are existing examples of light annotation tasks in the biomedical domain. The 2007 i2b2 NLP challenge task of identifying the smoking status of patients is a perfect example of a light annotation for a biomedical task, and one that will be discussed later in this paper. Similarly, the Automated Retrieval Console (ARC) system seems to be designed around the idea of asking only for light annotations from users.

The existence of these tasks proves that light annotation projects can be undertaken to yield datasets that represent complex information, but are themselves not complex, and can also later be useful for machine learning projects. However, so far no guidelines or methodology has been established for generalizing these types of tasks.

While investigating useful annotations in biomedical texts, Wilbur et al. (2006) identified five aspects of scientific papers that can be used generally in text mining: focus, polarity, certainty, evidence, and directionality. They reported that the inter-annotator agreements resulted in scores between 70 and 80 percent, which are good indications of an accessible and useful annotation task.

Another example of biomedical annotation is the semantic annotation done by Kim et al. (2008) over the GENIA corpus. In light of the complexity of the task, they employed what they call Single-Facet annotation, a system of presenting annotation tasks to the annotator in order to reduce the cognitive load on the annotators by "defining one aspect of the text as the focus of annotation". This is similar to the annotation approach used in the Brandeis Annotation Tool (BAT), which reduces error in an annotation project by reformulating an annotation task to be performed one layer at a time (Verhagen, 2010).

More generally, the fields of Corpus and Computational Linguistics have yielded specific criteria and methodologies for creating annotation tasks: the MATTER cycle provides a generalized system for developing annotated corpora (Pustejovsky, 2006), the Linguistic Annotation Framework (LAF) is part of an ISO standard for representing annotations in ways that ensure compatibility with other projects (Ide and Romary, 2006), and the seven maxims for annotation tasks identified by Leech (Leech, 1993) have been largely unchallenged over the years.

3. The MATTER Cycle

MATTER is a development cycle for natural language processing tasks involving annotation and machine learning. The steps are: *Model, Annotate, Train, Test, Evaluate, Revise* (Pustejovsky, 2006; Pustejovsky and Stubbs, 2012) (See Figure 1). MATTER represents a general methodology of standard development for all types of annotation tasks.

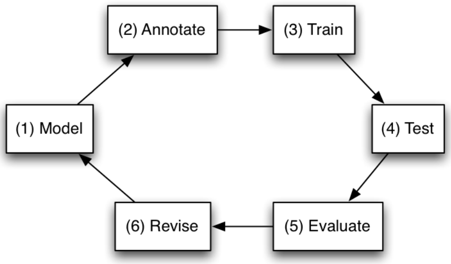


Figure 1: The MATTER cycle.

Within the MATTER cycle there is a smaller development cycle related specifically to the Model and Annotation phases—often when creating an annotation task, the model and annotation are re-evaluated and modified multiple times before the algorithm training is even attempted (see Figure 2). This is referred to as the MAMA (Model-Annotate-Model-Annotate), or the "babbling" phase of the development cycle, as it is the part of the process where the model and annotation become fully formed as a representation of the task (Pustejovsky and Moszkowicz, 2012).

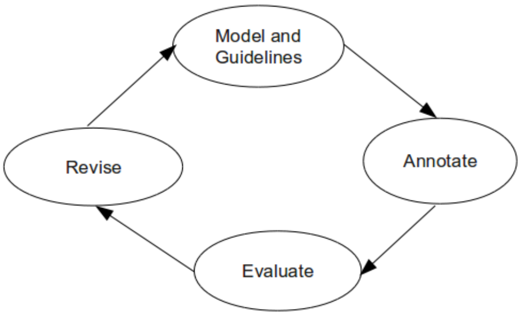


Figure 2: The Model-Annotate Cycle

For most annotation tasks, the Model is the specification used to describe the features of the annotation being applied to the corpus. It defines the tags, attributes, and metadata that will be represented by the annotators over the data being analyzed. The model, M, can be represented as a tuple: M = <T, R, I> where T is the vocabulary of terms, R is the relations between the terms, and I is their interpretation (Pustejovsky and Stubbs, 2012). In most tasks, a single model is used to represent all the information that is needed for the corpus being annotated; that is to say that, even in cases where a task may be divided into steps (for example, event tagging is done first, relation tagging is done later), the annotation output is a unified, complete representation of the desired model.

While this approach has worked well for many annotation tasks, it is easy to find examples of data mining questions where it would be impractical to ask a medical professional to provide all relevant aspects of annotation. Consider again the i2b2 challenge regarding smoking status (Uzuner et al., 2007). While the primary task—determining which of 5 categories (past smoker, current smoker, smoker, non-smoker, and unknown)—is straightforward, and can be represented simply as a single label applied to an entire document, there are many other factors involved in making that classification. In a document, problematic issues can include: which person is being described (some medical documents provide family history as well as patient history), the age of the document, the presence of negations, the scope of a modifying clause, ambiguities introduced by coreference, etc., not to mention linguistic features such as part-of-speech tags, chunking, tokenization, and so on. These features are not trivial to encode, but the results of the challenge show that at least some of them were used in top-performing systems (Clark et al., 2008; Cohen, 2008; Szarvas et al., 2006).

The Smoking Status dataset is an excellent example of a successful light annotation task specifically because it does *not* include any of those linguistic features. In the next section I will discuss what characteristics make it a good example of light annotation, and how those characteristics can be generalized to other annotation tasks.

4. Creating Light Annotation Tasks

The Smoking Status dataset has a number of attributes which individually may not make it a noteworthy task, but taken together provide an excellent example of a light annotation task. These are (Uzuner et al., 2007):

- The annotation was done by professionals in a field directly related to the area of study (pulmonologists);

- The task used only 5 categories of classification, and two of those contained degrees of uncertainty. For the specific classifications, Past Smoker was someone who quit a year or more ago, Current Smoker is someone who smoked within the past year, and a Non-Smoker is someone who never smoked. Less specifically, Smoker was used to classify someone who was either a current or past smoker, but the temporal reference was vague, and Unknown was used for files where no reference to smoking was made at all;

- The categories used were based on current medical practices and understanding. A layperson would be inclined to label someone who quit smoking 3 months ago as a past smoker, but someone in the medical profession would know that, because the effects of smoking are long-lasting, even people who recently quit are considered smokers for up to a year afterwards;

- Information about both textual and intuitive classifications were collected, though only textual information was used for the challenge due to disagreements over intuition.

Within this task there are some points that can be generalized to other tasks that intend to examine complex biomedical questions. By merging these points with existing linguistic annotation standards, we can establish maxims for creating good light annotation tasks. The following guidelines for creating light annotation tasks in the biomedical domain are therefore proposed:

- The annotations are performed by experts in the field;

- The task is divided into as few classification questions as possible;

- The classifications used in the model are based on current best biomedical theories and practices;

- Annotation should be done based only on what is in the text, not on expert's intuitions about the text.

- If possible, the annotations should be applied to sentence- or phrase-level sections of the document, or applied as labels to the entire text;

- Additional layers of annotation can be provided before or after the light annotation is performed without conflicting with the given classifications.

These guidelines are primarily targeted at projects that are looking to extract domain expert knowledge from texts (this paper focuses on biomedical examples, but this could be applied to other forms of expertise as well). That is, the Smoking Status was determined by pulmonologists because it is a subject which is directly related to their professional knowledge, and may not be easily interpreted by a layperson. Projects looking to add linguistic information, such as part-of-speech tags, to a text probably do not need to take this approach.

Let us examine each of these guidelines in turn:

Expert annotators: If the purpose of the task being performed is to learn something complex about the data, the annotations should be done by people who are qualified to make those determinations. On the surface this is obvious, but it is a departure from more traditional linguistic annotations, where linguists and doctors have shown roughly equal ability to apply part-of-speech tags, tree structures, and coreference markers (Tateisi and Tsujii, 2004; Tateisi et al., 2005; Cohen et al., 2010).

Minimal classifications: By breaking down the needed information into a small set of classification tasks (or even a single task, as is seen in the Smoking Status corpus), the annotation can be done much more quickly and accurately. This is particularly helpful for research groups who may not have a biomedical professional in house, but instead need to hire domain expert annotators as consultants: a process that can be costly and time-consuming. Wilbur et al. (2006) used a similar approach in their text classification task to great success. This is also similar to the Single-facet Annotation as explored by Kim et al. (2008).

Based on current theories and techniques: Beyond simply suggesting that annotations should not be intrinsically unscientific, the point of this guideline is to say that the medical or biomedical understanding of the text should take *precedence* over strictly linguistic analyses. For the Smoking Status corpus, for instance, a textual reading of 'quit smoking 3 months ago' by a layperson would indicate a status of 'Past Smoker', but that would be incorrect according to the medical interpretation. The annotation must thus reflect medical standards, and not be subordinated to easier or more obvious interpretations.

Evidence-based annotations: It seems reasonable to suggest that, if supplied with an expert's knowledge in a field, making use of the intuitions that go along with that knowledge would be a great boon to interpreting biomedical texts. However, both the Smoking Status challenge and Kim et al. found that leveraging expert knowledge resulted in greater discrepancies in inter-annotator agreement (Uzuner et al., 2007; Kim et al., 2008). Kim et al. relied instead on what they referred to as *Text-bound annotation*: annotations that required the annotators to “indicate clues in the text for every annotation they made”. This resulted in higher inter-annotator agreement and more useful annotations. There is a key difference between making use of expert *knowledge* and relying on expert *intuition*. Relying on intuition may result in annotators trying to read between the lines of a text, or past experience that tells them, ‘If a patient says this, it’s usually actually that’. Limiting annotations and classifications to what is said in the text will result in annotations that are both more agreed upon between annotators, and more useful for machine learning, if that is your goal.

Sentence- or phrase-level annotations: Once the annotation task has been cast as one of simple classification, it becomes much easier to instruct domain expert annotators to find sentences or phrases that are used to determine what classification a document or document section should be given. This task can be done much more quickly if the annotators are not asked to create careful markups of the entire document, but rather just to highlight the relevant portions, add a classification label, and then move on.

No conflict with additional annotations: This guideline applies to the practical matter of the actual encoding of the annotation. The annotation task should not rely on tools or outputs that will not be compatible with other layers of annotation. The easiest way to ensure this is to use tools that are LAF-compliant (Ide and Romary, 2006), and to represent annotations in stand-off XML or a similar scheme that does not change the text being annotated. This will make it easier to add layers of other annotations later in the process for use in machine learning.

Overall, the purpose of the light annotation task using this methodology is not necessarily to create a complete representation of all the relevant data in a biomedical text. It can, however, create a highly accurate layer of annotation that will be used in conjunction with other linguistic information, as was the case with the Smoking Status challenge. In terms of the MATTER cycle, the light annotation is not the full representation of the Model (M = <T, R, I>). Rather, the light annotation Model is a top-level set of annotation that is used to indicate portions of the document relevant to the classification, or to apply a label to a document as a whole. It does not represent the entire set of features necessary to create an algorithm (during the Training and Testing phases of MATTER) that is able to generate the desired classifications.

The light annotation can and should still be done in the context of the MATTER and MAMA cycles, as they represent established guidelines for text annotation tasks. The next section discusses a corpus of medical documents and examines how the MATTER and light annotation guidelines were applied to an annotation task using that data.

5. Case Study: Finding Patients who Match Selection Criteria

Finding patients who are eligible for participation in medical studies is not a trivial task, even when hospital billing codes can be used to help narrow the field of candidates. At some point medical records need to be examined, and that process is time-consuming and error-prone due to the complexity of the documents being reviewed.

Under a grant from the NIH (NIHR21LM009633-02, PI: James Pustejovsky), this problem was explored in collaboration with the Channing Laboratory at Brigham and Women’s Hospital and Harvard Medical.

In order to explore the possibility of automating at least part of the selection process, a test set of selection criteria for a mock case-control study was created, as well as a set of matching criteria in the interest of exploring the information required to create matched case-control groups. A set of 100 discharge summaries was selected from the MIMIC II Clinical Database (Clifford et al., 2010) for review. Documents were chosen based on keywords relevant to the chosen criteria; for a full discussion of the corpus selection process see Stubbs (forthcoming).

The complexities of representing eligibility criteria have been and are still being explored (Weng et al., 2010; Weng et al., 2011). However, rather than focusing on that aspect of the eligibility problem, this annotation effort looked specifically at what would be required for information extraction from the discharge summaries themselves.

The selection and matching criteria used to identify patients for the study were:

Selection criteria:
General criterion 1: must be under 55 years old at time of admission
General criterion 2: must have diabetes
Case criterion 1: must have had a cardiac event within 2 years of admission date
Control criterion 1: no history of cardiac events

Matching criteria:
Matching Criterion 1: race
Matching Criterion 2: sex
Matching Criterion 3: lipid measurement w/in 6 months of admission
Matching Criterion 4: information on diabetic treatment
Matching Criterion 5: lipid medications

It was immediately clear that a great deal of information would be required to automate this task with any degree of accuracy: document structure, temporal processing, and event recognition would likely be necessary, and possibly other information as well. However, given the complexity and domain-specific vocabulary of the discharge summaries, it was obvious that the document analysis would have to be done, at least in part, by someone working in medical research.

5.1. Annotation Task

Initially this project was going to use the Clinical E-Science Framework (CLEF) (Roberts et al., 2007; Roberts et al., 2008) annotation schema and guidelines (working group, 2007). CLEF has two extent annotations, *Entities* and *Signals*, and two link annotations, *Coreference* and *Relationships*. Each of these tags has subcategories that are used to further classify the text being annotated; for example, *Entities* is further subdivided into Condition, Intervention, Investigation, Result, Drug or Device, and Locus.

However, an initial annotation effort using only the different *Entities* tags quickly made it apparent that such an approach would be extremely time-consuming, and would also require substantial effort to be feasible. The existing CLEF guidelines were found to be unclear in terms of defining what made something a 'condition' rather than a 'result', or an 'intervention' instead of an 'investigation'. Unfortunately, the CLEF corpus is not available to the public and so it could not be used as a resource for making these distinctions.

While an underspecified annotation guideline is not an insurmountable problem, the deciding factor in moving away from CLEF was the amount of time it would take to perform the annotation. For each document that was annotated by a Registered Nurse (i.e., someone who was familiar with the terminology and structure of the files in the corpus), annotating only the entities took several hours. It was clear that the budget for the grant could not support such an intensive annotation project, and a different system would have to be used.

Therefore, it was agreed that the document annotation would be broken down into parts: the linguistic processing (part-of-speech, temporal processing, dependency parsing) could be done by the computational linguistics researchers as needed, while the determination of who met what criteria would be completed as a light annotation task by the medical researchers.

The annotation was done by two medical researchers: one is a Registered Nurse, and the other is involved in patient selection and data collection for medical studies. Because the discharge summaries being examined were so dense with information, rather than have the annotators give a single label per criterion to each document, they were asked to indicate which parts of the document were relevant to each criteria.

The annotation scheme used only four tags: three extent tags used to identify sections of text relevant to each criterion, and one linking tag used to associate different extents where necessary. More specifically:

The **selection_criterion** and **matching_criterion** tags were used to mark text that was relevant to the criteria described above. Both **criterion** tags have an attribute called "criterion", which annotators used to indicate which criterion the text they were marking was relevant to, and another attribute was used to indicate whether the annotated text showed that the criterion was met or not (or present or not, in the case of matching criteria).

The **modifier** tag was used to annotate contexts (such as adjectival phrases) that would change the interpretation of the criterion-related text. The use of this tag varied widely: in some cases it was used to mark dates related to time-dependent criteria, in others it was used to indicate if the criterion-related text was about a family member rather than the patient, or was in some way negated or theorized about (e.g., "may be at risk for..."). In order to create a connection between criterion-related text and the modifying extents, a **modifies** link tag was used to connect the two spans where needed.

For the phrase "father with DMII" the resulting annotation would look like this:

```
<Selection_criterion id="SC16"
   text="DMII" criterion="diabetes"
   meets="NO" />
<Modifier id="M2" text="father" />
<Modifies id="ML26" from="M2"
   to="SC16"/>
```

The annotators did not have to give a document-level classification to each discharge summary; rather, the status of each file in relation to the established criteria was determined automatically after adjudication was performed. Annotations were done in MAE (Multi-purpose Annotation Environment), a intuitive light-weight annotation tool that did not require the annotators to be trained in using a complex software package or understand the underlying XML representation (Stubbs, 2011).

Because the annotators were asked to mark only the parts of the document that were relevant to the criteria, the annotation process was able to go much faster than when the CLEF annotation was attempted. Using CLEF, a single document took hours for the annotator to generate, while with this scheme an average of three documents an hour could be marked up.

This annotation scheme adheres to the guidelines for light annotation outlined above: the annotators had expert knowledge of the field; the task was reduced to a small set of classification tasks at the phrase level; the classifications were based on selection criteria modeled after existing studies; the use of the **modifier** tag required them to provide support for their claims when needed; and the annotations were encoded in stand-off XML so they can be distributed separately from the discharge summaries themselves, and later merged with other annotation layers.

5.2. Annotation Results

The annotation effort over the discharge summaries benefited from the concepts outlined in the guidelines for light annotation tasks. From the outset, the light annotation was much faster for the annotators to complete than the CLEF annotation: under CLEF, each document took roughly 2 hours to annotate, while under the light annotation an average of 3.72 documents could be annotated per hour. Because of the time saved by using the light annotation scheme, the actual cost of the annotation project was roughly 90% less than the projected cost of the CLEF annotation. This improvement in speed reflected the reduced cognitive load that the task placed on the annotators; both annotators felt that the light annotation task was much more tractable and easy to both understand and perform. By using the light annotation task, the data was encoded with expert opinions much more quickly and cheaply than if we had continued to use CLEF. Additionally, the format of the annotation is such that the textual evidence for their opinions can be analyzed later without consulting the annotators, and the format of the annotation is compatible with a variety of existing tools.

Aannotation tasks, however, are always iterative, and the results reported on here are from only the second iteration of the MAMA cycle over this data. As expected under the MATTER and MAMA methodologies, the analysis presented here revealed some problems with the model that can be corrected in later annotations. Specifically, the level at which the annotators were asked to evaluate the text will be expanded—rather than use the **modifier** and **modifies** tags, only the **criterion** tags will be used, and the annotators would be instructed to annotate the entire section relevant to the criteria, including any modifying phrases. This would both cut down on the confusion over how to use the **modifier** tag, as well as speed up the annotation process even more. In future work, the definition of 'cardiac event' should also be more clearly defined in the guidelines.

Due to the complexity of the discharge summaries being annotated, agreement scores based on extent markup between the two annotators were not as high as can be achieved in later iterations. They both generally agreed on what aspects of the text were relevant to the criteria, and which patients met and did not meet the different requirements. However, because of the density of the texts and the amount of repetition in each record, the exact extents that they used to make those determinations did not always overlap, though they were often complementary. For example, Annotator 1 may have spotted a mention of "type2dm+" in the "patient medical history" section of the record, but missed the "patient has diabetes" phrase in the "hospital course" portion of the document, while Annotator 2 did the opposite. Therefore, while Cohen's Kappa (Cohen, 1960) for all the extents marked by each annotator is .505, when compared to the Gold Standard corpus each annotator had high precision (an average of .92) and lower recall (an average of .84) for the **criterion** tags, indicating that both annotators had a higher percentage of false negatives, an analysis that backs up the interpretation of discharge summaries being particularly difficult to read closely for content.

The task of applying the Gold Standard to machine learning algorithms is still in progress, though it is benefiting greatly from the research done for SecTag (Denny et al., 2008; Denny et al., 2009) and cTAKES (Savova et al., 2010), and the temporal analyses of discharge summaries done by Hripcsak et al. (Hripcsak et al., 2009), Zhou et al.(Zhou et al., 2007) and Mowery et al. (Mowery et al., 2009).

6. Conclusions and Future work

This paper presents a methodology for creating light annotation tasks, specifically in the biomedical domain. The guidelines presented represent suggestions extracted from other recognized endeavors, and are based on solid theoretical and practical foundations. While a full application of the light annotation methodology to the case-control annotation task has not yet been completed, it is clear that even these preliminary results show improvements over what would have been achieved with a 'heavier', more complex annotation task.

Using this methodology for extracting light annotation tasks from more complicated endeavors is a viable way for researchers who want to process biomedical texts to approach the problem, without being expert knowledge of the chosen fields themselves. These guidelines enable researchers to obtain contentful, evidence-based expert analyses of domain-specific texts without excessive cost or time investments. This approach could also be used in conjunction with other systems that have been designed for enhancing annotation systems, such as an accelerated annotation framework (Tsuruoka et al., 2008).

While not all tasks are necessarily going to be able to be converted into this format, those that are may benefit from using this approach, particularly in labs where access to domain experts is limited. Admittedly, there is potential for data loss in using a light annotation framework—if the task is not sufficiently well-defined in relation to the goal of the annotation task, there is potential for wasted effort. While this is true of any annotation task, because the focus of this effort is to assist programs where access to domain experts is limited and therefore more costly, it is imperative that any light annotation task undertaken with limited resources be carefully considered in terms of utility.

The guidelines presented here are not limited to use with annotation efforts in the biomedical domain; they can be used for any light annotation task requiring expert knowledge. Applying the methods described here to other domains should be explored for future research.

7. Acknowledgements

I would like to BJ Harshfield, Cheryl Keenan, and Vincent Carey for their assistance with the annotation task, and my advisor James Pustejovsky for his help clarifying the theories presented here. Thanks also to Lotus Goldberg, Jessica Moszkowicz, and Marc Verhagen for proofreading and providing comments on this paper.

Funding for this research was provided by NIH grant NIHR21LM009633-02, PI: James Pustejovsky.

8. References

Cheryl Clark, Kathleen Good, Lesley Jezierny, Melissa Macpherson, Brian Wilson, and Urszula Chajewska. 2008. Identifying smokers with a medical extraction system. *Journal of American Medical Informatics Association*, 15:36–39.

G. Clifford, D. Scott, and M. Villarroel. 2010. User guide and documentation for the mimic ii database. http://mimic.physionet.org/UserGuide/UserGuide.html, August. accessed Nov. 23, 2010.

K. Bretonnel Cohen, Arrick Lanfranchi, William Corvey, William A. Baumgartner Jr., Christophe Roeder, Philip V. Ogren, Martha Palmer, and Lawrence Hunter. 2010. Annotation of all coreference in biomedical text: Guideline selection and adaptation. *BioTxtM 2010: 2nd workshop on building and evaluating resources for biomedical text mining*, pages 37–41.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Aaron M. Cohen. 2008. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *Journal of American Medical Informatics Association*, 15:32–35.

Leonard W D'Avolio, Thien M Nguyen, Wildon R Farwell, Yongming Chen, Felicia Fitzmeyer, Owen M Harris, and

Louis D Fiore. 2010. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (arc). *J Am Med Inform Assoc*, 17(4):375–82.

Leonard W D'Avolio, Thien M Nguyen, Sergey Goryachev, and Louis D Fiore. 2011. Automated concept-level information extraction to reduce the need for custom software and rules development. *J Am Med Inform Assoc*, 18(5):607–13.

Joshua C Denny, Randolph A Miller, Kevin B Johnson, and Anderson Spickard. 2008. Development and evaluation of a clinical note section header terminology. *AMIA Annual Symposium proceedings*, pages 156–60.

Joshua C Denny, Anderson Spickard, Kevin B Johnson, Neeraja B Peterson, Josh F Peterson, and Randolph A Miller. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc*, 16(6):806–15.

George Hripcsak, Noémie Elhadad, Yueh-Hsia Chen, Li Zhou, and Frances P Morrison. 2009. Using empiric semantic correlation to interpret temporal assertions in clinical texts. *J Am Med Inform Assoc*, 16(2):220–7.

Nancy Ide and Laurent Romary. 2006. Representing linguistic corpora and their annotations. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*.

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

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 Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Junichi Tsujii. 2011. Extracting biomolecular events from literaturethe bionlp09 shared task. *Computational Intelligence*, 27(4):513–540.

Geoffrey Leech. 1993. Corpus annotation schemes. *Lit Linguist Computing*, 8(4):275–281.

Danielle L. Mowery, Henk Harkema, John N. Dowling, Jonathan L. Lustgarten, and Wendy W. Chapman. 2009. Distinguishing historical from current problems in clinical reports: which textual features help? In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 10–18, Morristown, NJ, USA. Association for Computational Linguistics.

James Pustejovsky and Jessica L. Moszkowicz. 2012. The role of model testing in standards development: The case of iso-space. In *In the Proceedings of the Eight International Conference on Language Resources and Evalutation (LREC)*.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning.* O'Reilly Media Inc.

James Pustejovsky. 2006. Unifying linguistic annotations: A timeml case study. In *Proceedings of Text, Speech, and Dialogue Conference*.

Angus Roberts, Robert Gaizauskas, and Mark et al Hepple. 2007. The clef corpus: semantic annotation of clinical text. *AMIA Annual Symposium proceedings*, pages 625–9.

A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, and A. Setzer. 2008. Semantic annotation of clinical text: The clef corpus. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining.*

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–13.

Amber Stubbs. 2011. Mae and mai: Lightweight annotation and adjudication tools. In *Proceedings of the Linguistic Annotation Workshop*, Portland, OR.

Amber Stubbs. forthcoming. *A Methodology for Leveraging Professional Knowledge in Corpus Annotation.* Ph.D. thesis, Brandeis University.

G. Szarvas, R. Farkas, S. Ivn, A. Kocsor, and R. Busa Fekete. 2006. Automatic extraction of semantic content from medical discharge records. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data.*

Yuka Tateisi and Jun'ichi Tsujii. 2004. Part-of-speech annotation of biology research abstracts. In *Proceedings of the 4th International Conference on Language Resource and Evaluation.*

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotationfor the genia corpus. In *Proceedings of the IJCNLP, companion volume*.

Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. 2008. Accelerating the annotation of sparse named entities by dynamic sentence selection. *BMC Bioinformatics*, 9.

Ozlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2007. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, 15(1):14–24.

Marc Verhagen. 2010. The brandeis annotation tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Chunhua Weng, Samson W Tu, Ida Sim, and Rachel Richesson. 2010. Formal representation of eligibility criteria: a literature review. *J Biomed Inform*, 43(3):451–67.

Chunhua Weng, Zhihui Luo, and Steven B. Johnson. 2011. Elixr: An approach to eligibility criteria extraction and representations. In *2011 CRI Summit Proceedings*. accessed Nov. 22, 2010.

W. John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC B*, 7.

CLEF working group, 2007. *CLEF Annotation Guidelines*. `http://nlp.shef.ac.uk/clef/TheGuidelines/TheGuidelines.html`, May. accessed March 2010.

Li Zhou, Simon Parsons, and George Hripcsak. 2007. The evaluation of a temporal reasoning system in processing clinical discharge summaries. *J Am Med Inform Assoc*, 15(1):99–106.