

# INFORMATION EXTRACTION FROM BIOMEDICAL TEXT: THE BIOTEXT PROJECT

**Filip Ginter, Tapio Pahikkala, Sampo Pyysalo, Evgeni Tsivtsivadze  
Jorma Boberg, Jouni Järvinen, Aleksandr Mylläri and Tapio Salakoski**  
Turku Centre for Computer Science (TUUS) and Dept. of IT, University of Turku

## Abstract

We study information extraction for identifying protein-protein interactions stated in biomedical text. In this paper, we present an architecture for an information extraction system and discuss our improvements and results pertaining to several components of the system, including information retrieval, named entity recognition, syntactic analysis, and domain analysis. The individual results are discussed in the context of the whole system, and domain adaptations and differences from classical approaches are considered. We combine structural natural language processing with machine learning methods to address the general and domain-specific challenges of information extraction targeting protein-protein interactions.

**Keywords:** biomedical literature mining, information retrieval, named entity recognition, word sense disambiguation, parsing, parse ranking

## 1. Introduction

The amount of published knowledge in the biomedical domain is overwhelming and grows at an unprecedented rate. Although many databases collecting biomedical knowledge exist, their coverage is limited and manual identification of e.g. protein-protein interactions requires significant human effort. Freeform text remains a main source of information and thus Natural Language Processing (NLP) and Information Extraction (IE) methods are required to facilitate automated processing and structured access to the knowledge. The BioText project aims at developing NLP methods and resources for biomedical text mining as well as adapting existing methods to take into account the specific properties of the biomedical text domain. This paper gives an overview of our approach, the developed methods and the key results of the project.

Our overall goal is the development of a modular system that processes biomedical text, such as abstracts contained in the PubMed literature database, and extracts the protein-protein interactions stated therein. The system consists of the following major subsystems: Information Retrieval (IR), Named Entity (NE) recognition, syntactic analysis, and pattern-based domain analysis. We apply machine learning approaches such as Bayesian classification, Support Vector Machines (SVM) (see e.g. Vapnik 1998) and Regularized Least-Squares (RLS) (see e.g. Poggio and Smale 2003) as well as

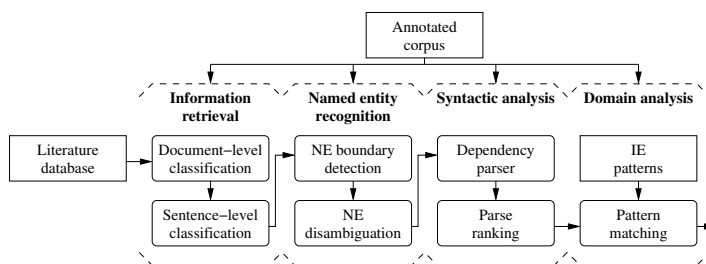


Figure 1: The IE system architecture

structural linguistic methods such as dependency-based syntactic analysis. We also develop methods that combine the two general approaches, taking advantage of both explicit linguistic knowledge and machine learning. For a recent thorough review of related work in Bio-NLP, see for example Cohen and Hunter (2004).

The architecture of our IE system is illustrated in Figure 1. The following sections describe the components of the system in detail.

## 2. Annotated domain-language corpus

An annotated domain-language corpus is necessary to facilitate the development and evaluation of the various parts of IE systems. We have created a corpus of biomedical English focused on protein-protein interactions. The corpus consists of 1100 sentences manually annotated at three levels: NEs, dependency syntax, and entity interactions. It can thus provide data for the development and evaluation of all of the key components of the IE system. Further, as all the levels of annotation are provided for a single set of sentences, the corpus allows the components of the IE system to be tested not only individually but also as an integrated whole. The corpus is described in detail in Ginter et al. (2004d) and will be made publicly available at <http://www.cs.utu.fi/bdb>.

## 3. Information retrieval

Many of the steps in the IE system, e.g. the full syntactic analysis, are computationally costly. Fully processing a large literature database such as PubMed, which contains 7.5 million article abstracts, is thus not practical. We therefore study IR methods that retrieve from publications only the sentences which are relevant to the domain of interest. While many standard approaches to IR have been described in literature, we study methods that utilize information specific to the biomedical domain. In Ginter et al. (2004c), we introduced a method applicable to the classification of PubMed-indexed articles. We devised a scheme for transforming the MeSH biomedical ontology used to index PubMed articles, and showed that the ontology transformations lead to an increase in classification performance. To identify individual sentences likely to discuss protein-protein interactions, we also introduced a method in which known protein names, verbs specific to protein-protein interactions, and their mutual positions in the sentence are used as features for a rough-set based classifier (Ginter et al. 2004b).

## 4. Named entity recognition and disambiguation

NE recognition can be divided into two subtasks: determining the boundaries of the NEs and classifying the entities into classes such as genes and proteins. Both problems can be addressed using Word Sense Disambiguation (WSD) methods. Much of the ambiguity in biomedical text is caused by inconsistent or non-existent naming conventions. Further, capitalization and other surface clues are not reliable indicators of entities in the domain. For example, there exist *Drosophila* gene names such as *white* and *cycle* which can be confused with the ordinary meanings of these words. We use machine learning methods with particular focus on kernel-based learning algorithms (see e.g. Schölkopf and Smola 2002) to address the problem of WSD.

In Ginter et al. (2004a), we introduced a statistical classification method and a weighted bag-of-words representation, where the context words are weighted so that the words located closer to the ambiguous word receive higher weights. The new method was shown to improve the classification performance in gene/protein name disambiguation from 79% to 82% accuracy.

We have adapted the weighted bag-of-words approach for SVM classifiers and applied them to the problem of gene/protein name disambiguation, improving the performance, measured as the area under the ROC curve (AUC), from 80% to 85% (?). We have also introduced a position-sensitive kernel function which generalizes over the ordinary bag-of-words, position-sensitive bag-of-words and weighted bag-of-words approaches (Pahikkala et al. 2005b). Considering context-sensitive spelling error correction as a WSD problem, it was demonstrated that the position-sensitive kernel improves the performance of the SVM classifier from 94% to 98% (AUC). The results reflect the difficulty of the biomedical disambiguation tasks as well as demonstrate the applicability of the method to other domains.

In Pahikkala et al. (2005a), we further analyze this kernel function and construct smoothed word position-sensitive as well as smoothed word position- and distance-sensitive representations of our training data using kernel density estimation techniques (see e.g. Silverman 1986). For the Naïve Bayes classifier, these representations were used to obtain class-conditional probabilities of word-position features. We demonstrate with the Senseval-3 data that the kernel improves the classification performance of SVMs compared to the ordinary Bag-of-Words kernel and furthermore improves the classification performance of the Bayes classifier given the kernel-smoothed data representation.

## 5. Syntactic analysis

In this section, we present our choice of parser and the architecture of the syntactic analysis component. We also illustrate our use of machine-learning methods to improve the performance of a parser based on a hand-written grammar.

### 5.1. Parser

Our analyses suggest that general English parsers may not be well applicable to biomedical English, and that adaption to the domain is required (Pyysalo et al. 2004: 2005). Moreover, a statistical inference of the domain grammar is infeasible as the amount of treebank data in the domain is very limited—the largest domain treebank is the GENIA

treebank<sup>1</sup> with 1700 sentences. This motivates the choice of a parser based on a hand-written grammar that can be manually adapted to the domain; in the BioText project, we have decided to use the Link Grammar (LG) parser of Sleator and Temperley (1991). The LG parser is a full dependency parser with broad coverage of newswire English. LG has recently received significant attention in the Bio-NLP domain, see for example Alphonse et al. (2004).

The architecture of our syntactic analysis component built around LG is as follows. First, the input sentences are tokenized in a separate tokenization step: The tokenization model originally used by LG was found unsuitable for many common features of biomedical text and replaced with an external tokenization system. After tokenization, we have chosen to augment the parsing system with separate preprocessing and postprocessing stages. In preprocessing, input sentences are simplified by replacing detected NEs with single tokens recognized by the parser, as well as by removing citations and other features for which the parser has no support and which can be naturally captured using regular expressions. Postprocessing is applied after parsing to restore the original sentence text. To improve the applicability of LG to the biomedical domain, we have implemented a number of the modifications proposed in Pyysalo et al. (2004). While the work on parser adaptation is still undergoing, preliminary evaluation suggests that the implemented modifications increase the fraction of recovered correct dependencies from 73% to 78% in the parse ranked first by the built-in heuristics of the LG parser. The parser generates all the alternative parses allowed by the grammar; the respective improvement for the best generated parse is from 82% to 89%.

The domain analysis is performed on the first parse returned by the syntactic analysis component. We have found that the heuristic parse ranking of LG often performs poorly, failing to rank the best parses first. To address this issue, we are developing a machine-learning approach for parse ranking that is applied after post-processing.

## 5.2. Parse ranking

The task of recognizing the best parses among a set of alternative parses for a single sentence can be cast as a ranking problem. We are currently developing a ranking machine based on the RLS algorithm. Our methodology couples RLS, different ranking performance measures and grammatically motivated features. To convey the most important information about parse structure to the ranking machine, we apply features such as grammatical bigrams, link types (the grammatical roles assigned to the links), a combination of link length and link type, part-of-speech information, and several additional attributes. Each parse is assigned a penalty based on the number of incorrect links. We are also studying a scoring approach where additional information about link types is used in penalization.

The developed method, Regularized Least-Squares Ranking (RLSRa), is a special case of ordinal regression where performance evaluation is based on the rank correlation measure of Kendall (1970), scaled between zero and one. Preliminary evaluation of RLSRa against LG parser built-in heuristics indicates a performance improvement from 55% to 70% using our method. Furthermore, RLSRa provides a reliable ranking solution in application to sparse biomedical datasets.

---

<sup>1</sup><http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/GTB.html>

## 6. Domain analysis

To extract factual knowledge from the parsed sentences, we are developing a set of hand-written patterns. Each pattern specifies a substructure of the linkage (the graph that represents an LG dependency parse) that is likely to state a protein-protein interaction. A successful match of a pattern in a linkage corresponds to an identified interaction. We chose to create high-precision patterns to minimize the number of false positive matches. High precision typically implies low recall; however, due to the large amount of published literature, the system can be given more than one opportunity to extract most of the interactions as they are likely to be stated in several publications. Processing more data can thus diminish the low recall problem to some extent.

The choice of parser has an obvious influence on the nature of the patterns and the formalism in which the patterns are expressed. Since we chose a full dependency parser, it is natural to represent both the linkage and the patterns in terms of relations on the set of words and link types. This representation is naturally and straightforwardly expressed in a declarative language such as Prolog. Each linkage and each pattern are thus described as a set of predicates, and the unification mechanism of Prolog provides the pattern matching mechanism.

## 7. Conclusions and future work

We have described our work in biomedical IE, presented the architecture of an IE system targeting protein-protein interactions, and discussed each of its components. For each part of the system, we have presented our approach, summarizing improvements and key results. Currently, we are focusing on finishing the domain adaptation of the syntactic analysis component and the development of IE patterns. The implementation of an integrated system that combines the discussed components remains future work.

## Acknowledgements

This work has been supported by Tekes, the Finnish National Technology Agency.

## References

- Alphonse, Erick; Aubin, Sophie; Bessi eres, Philippe; Bisson, Gilles; Hamon, Thierry; Lagarrigue, Sandrine; Nazarenko, Adeline; Manine, Alain-Pierre; N edellec, Claire; Vetah, Mohamed Ould Abdel; Poibeau, Thierry; Weissenbacher, Davy 2004. Event-Based Information Extraction for the biomedical domain: the Caderige project. In: *Proceedings of the JNLPBA workshop at COLING'04, Geneva*. 43–49
- Audibert, Laurent 2004. Word sense disambiguation criteria: a systematic study. In: *Proceedings of COLING'04, Geneva*, Association for Computational Linguistics. 910–916
- Cohen, K. Bretonnel; Hunter, Lawrence 2004. In: Dubitzky, Werner; Pereira, F. (eds.), *Artificial intelligence and systems biology*, Kluwer Academic Publishers
- Ginter, Filip; Boberg, Jorma; J arvinen, Jouni; Salakoski, Tapio 2004a. New techniques for disambiguation in natural language and their application to biological text. In: *Journal of Machine Learning Research* 5, 605–621

- Ginter, Filip; Pahikkala, Tapio; Pyysalo, Sampo; Boberg, Jorma; Järvinen, Jouni; Salakoski, Tapio 2004b. Extracting protein-protein interaction sentences by applying rough set data analysis. In: *Proceedings of RSCTC'04*: Vol. 3066 of *Lecture Notes in Artificial Intelligence*, Springer, Heidelberg. 780–785
- Ginter, Filip; Pyysalo, Sampo; Boberg, Jorma; Järvinen, Jouni; Salakoski, Tapio 2004c. Ontology-based feature transformations: A data-driven approach. In: *Proceedings of EsTAL'04*: Vol. 3230 of *Lecture Notes in Artificial Intelligence*, Springer, Heidelberg. 279–290
- Ginter, Filip; Pyysalo, Sampo; Heimonen, Juho; Boberg, Jorma; Järvinen, Jouni; Salakoski, Tapio 2004d. Bio Dependency Bank: a dependency corpus for information extraction in the biomedical domain. Submitted.
- Kendall, Maurice G. 1970. Rank Correlation Methods. Griffin, London
- Pahikkala, Tapio; Ginter, Filip; Boberg, Jorma; Järvinen, Jouni; Salakoski, Tapio 2004. Contextual weighting for support vector machines in literature mining: an application to gene versus protein name disambiguation. Submitted.
- Pahikkala, Tapio; Pyysalo, Sampo; Boberg, Jorma; Mylläri, Aleksandr; Salakoski, Tapio 2005a. Improving the performance of Bayesian and support vector classifiers in word sense disambiguation using positional information. Submitted.
- Pahikkala, Tapio; Pyysalo, Sampo; Ginter, Filip; Boberg, Jorma; Järvinen, Jouni; Salakoski, Tapio 2005b. Kernels incorporating word positional information in natural language disambiguation tasks. In: *Proceedings of FLAIRS'05*, Clearwater Beach, Florida. To appear.
- Poggio, T.; Smale, S. 2003. The mathematics of learning: Dealing with data. In: *Amer. Math. Soc. Notice* **50(5)**, 537–544
- Pyysalo, Sampo; Ginter, Filip; Pahikkala, Tapio; Boberg, Jorma; Järvinen, Jouni; Salakoski, Tapio 2005. Analysis of two dependency parsers on biomedical corpus targeted at protein-protein interactions. Submitted.
- Pyysalo, Sampo; Ginter, Filip; Pahikkala, Tapio; Boberg, Jorma; Järvinen, Jouni; Salakoski, Tapio; Koivula, Jeppe 2004. Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In: *Proceedings of the JNLPBA workshop at COLING'04, Geneva*. 15–21
- Schölkopf, Bernhard; Smola, Alexander J. 2002. Learning with kernels. MIT Press, Cambridge
- Silverman, B. W. 1986. Density estimation for statistics and data analysis. Chapman & Hall, London
- Sleator, Daniel D.; Temperley, Davy 1991. Parsing English with a link grammar. Technical Report CMU-CS-91-196: Department of Computer Science, Carnegie Mellon University, Pittsburgh
- Tsivtsivadze, Evgeni; Pahikkala, Tapio; Pyysalo, Sampo; Boberg, Jorma; Mylläri, Aleksandr; Salakoski, Tapio 2005. Regularized least-squares ranking for parse selection. Submitted.
- Vapnik, Vladimir 1998. Statistical Learning Theory. Wiley, New York

F. GINTER, T. PAHIKKALA, S. PYYSALO, AND E. TSIVTSIVADZE are postgraduate students at Turku Centre for Computer Science (TUCS), holding MSc. degrees in computer science and

working full-time for the BioText project since 2001, 2002, 2003, and 2004, respectively. E-mail: [firstname.lastname@it.utu.fi](mailto:firstname.lastname@it.utu.fi).

J. BOBERG, J. JÄRVINEN, A. MYLLÄRI, AND T. SALAKOSKI are lecturers at the Dept. of IT, University of Turku, holding PhD. degrees in computers science, mathematics, mathematics, and computer science, respectively. T. Salakoski is further a professor of computer science and vice-head of the department. E-mail: [firstname.lastname@it.utu.fi](mailto:firstname.lastname@it.utu.fi).