Sampo Pyysalo

A Dependency Parsing Approach to Biomedical Text Mining

Turku Centre *for* Computer Science

TUCS Dissertations
No 105, August 2008

A Dependency Parsing Approach to Biomedical Text Mining

Sampo Pyysalo

To be presented, with the permission of the Faculty of Mathematics and Natural Sciences of the University of Turku, for public criticism on September 6, 2008, at 12 o'clock.

University of Turku
Department of Information Technology
Joukahaisenkatu 3–5, 20520 Turku

2008

Supervisor

Professor Tapio Salakoski
Department of Information Technology
University of Turku
Finland

Reviewers

Dr Sophia Ananiadou
School of Computer Science,
University of Manchester
National Centre for Text Mining (NaCTeM)
United Kingdom

Dr Dietrich Rebholz-Schuhmann
Rebholz Group (Text Mining)
European Bioinformatics Institute
United Kingdom

Opponent

Professor Jun'ichi Tsujii
Department of Computer Science
University of Tokyo
Japan
School of Computer Science
University of Manchester
United Kingdom

ISBN 978-952-12-2131-6
ISSN 1239-1883

Abstract

Biomedical research is currently facing a new type of challenge: an excess of information, both in terms of raw data from experiments and in the number of scientific publications describing their results. Mirroring the focus on data mining techniques to address the issues of structured data, there has recently been great interest in the development and application of text mining techniques to make more effective use of the knowledge contained in biomedical scientific publications, accessible only in the form of natural human language.

This thesis describes research done in the broader scope of projects aiming to develop methods, tools and techniques for text mining tasks in general and for the biomedical domain in particular. The work described here involves more specifically the goal of extracting information from statements concerning relations of biomedical entities, such as protein-protein interactions. The approach taken is one using full parsing—syntactic analysis of the entire structure of sentences—and machine learning, aiming to develop reliable methods that can further be generalized to apply also to other domains.

The five papers at the core of this thesis describe research on a number of distinct but related topics in text mining. In the first of these studies, we assessed the applicability of two popular general English parsers to biomedical text mining and, finding their performance limited, identified several specific challenges to accurate parsing of domain text. In a follow-up study focusing on parsing issues related to specialized domain terminology, we evaluated three lexical adaptation methods. We found that the accurate resolution of unknown words can considerably improve parsing performance and introduced a domain-adapted parser that reduced the error rate of the original by 10% while also roughly halving parsing time.

To establish the relative merits of parsers that differ in the applied formalisms and the representation given to their syntactic analyses, we have also developed evaluation methodology, considering different approaches to establishing comparable dependency-based evaluation results. We introduced a methodology for creating highly accurate conversions between different parse representations, demonstrating the feasibility of unification of

diverse syntactic schemes under a shared, application-oriented representation. In addition to allowing formalism-neutral evaluation, we argue that such unification can also increase the value of parsers for domain text mining. As a further step in this direction, we analysed the characteristics of publicly available biomedical corpora annotated for protein-protein interactions and created tools for converting them into a shared form, thus contributing also to the unification of text mining resources. The introduced unified corpora allowed us to perform a task-oriented comparative evaluation of biomedical text mining corpora. This evaluation established clear limits on the comparability of results for text mining methods evaluated on different resources, prompting further efforts toward standardization.

To support this and other research, we have also designed and annotated BioInfer, the first domain corpus of its size combining annotation of syntax and biomedical entities with a detailed annotation of their relationships. The corpus represents a major design and development effort of the research group, with manual annotation that identifies over 6000 entities, 2500 relationships and 28,000 syntactic dependencies in 1100 sentences. In addition to combining these key annotations for a single set of sentences, BioInfer was also the first domain resource to introduce a representation of entity relations that is supported by ontologies and able to capture complex, structured relationships.

Part I of this thesis presents a summary of this research in the broader context of a text mining system, and Part II contains reprints of the five included publications.

Acknowledgments

I would first like to thank my supervisor, Tapio Salakoski, for his always good advice and for his patience, support and trust in allowing the pursuit of often distant and unsure goals in research. I am similarly indebted to my advisors Jorma Boberg and Jouni Järvinen, whose considerable efforts to enforce rigor, precision and statistical significance in my work have not been forgotten even as the focus of my study has increasingly moved toward language-related topics. None of the present research would have come to be without their vision and efforts. A special thank you also to Olli Mertanen, whose flexibility and understanding made the last year of this work possible.

The group in which I have had the privilege of working has been a constant source of inspiration, having a fruitful combination of people with very different but compatible tempers and interests. My gratitude goes out in particular to Filip Ginter, who has contributed to our shared goals his ideas, time and mental health far beyond the call of duty. The five years of our collaboration have shaped every aspect of this research. My thanks also, in alphabetical order for lack of a fair ranking criterion, to Antti Airola, Jari Björne, Katri Haverinen, Juho Heimonen, Veronika Laippala, Alexandr Mylläri, Tapio Pahikkala, Hanna Suominen and Evgeni Tsivtsivadze of TUCS and the University of Turku and to my gracious hosts and coauthors Sophie Aubin and Adeline Nazarenko of Université Paris-Nord for their many direct contributions to my research, their company, and the many stimulating discussions we have shared.

I am grateful to the reviewers of this thesis, Dr. Sophia Ananiadou and Dr. Dietrich Rebholz-Schuhmann, for their encouragement and careful criticism of my work. Their comments have strengthened both this thesis and my confidence to pursue these goals further. My thanks also to Professor Jun'ichi Tsujii for accepting to act as my opponent.

For making this research possible I owe a great debt—though fortunately not financial—to the many fine institutions that have supported it: first and foremost TUCS for their generous long-term commitment, and also Nokia, Suomalais-Ranskalainen Teknillistieteellinen Seura, and, through the projects I have worked in, Tekes – the Finnish Funding Agency for Technology and Innovation, and the Academy of Finland.

Most importantly, my thanks, which my command of English (or, indeed, of any language) is entirely insufficient to express, to my parents for their support and encouragement and to my wife Heidi for her love and her tolerance of extreme degrees of both my presence and my absence during this work.

List of original publications included in the thesis

I Pyysalo, S., Ginter, F., Pahikkala, T., Boberg, J., Järvinen, J., and Salakoski, T. (2006). Evaluation of two dependency parsers on biomedical corpus targeted at protein-protein interactions. *International Journal of Medical Informatics*, 75(6):430–442.

II Pyysalo, S., Aubin, S., Nazarenko, A., and Salakoski, T. (2006). Lexical adaptation of Link Grammar to the biomedical sublanguage: A comparative evaluation of three approaches. *BMC Bioinformatics*, 7(Suppl. 3):S2.

III Pyysalo, S., Ginter, F., Laippala, V., Haverinen, K., Heimonen, J., and Salakoski, T. (2007). On the unification of syntactic annotations under the stanford dependency scheme: A case study on BioInfer and GENIA. In *ACL'07 workshop on Biological, translational, and clinical language processing (BioNLP'07)*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

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

V Pyysalo, S., Airola, A., Heimonen, J., Björne, J., Ginter, F., and Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl. 3):S6.

List of related publications not included in the thesis

Co-authored publications and book chapters

- Airola, A., Pyysalo, S., Björne, J., , Pahikkala, T., Ginter, F., and Salakoski, T. (2008). A graph kernel for protein-protein interaction extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP'08)*, pages 1–9. Association for Computational Linguistics.

- Björne, J., Pyysalo, S., Ginter, F., and Salakoski, T. (2008). How complex are complex protein-protein interactions? In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

- Ginter, F., Pahikkala, T., Pyysalo, S., Boberg, J., Järvinen, J., and Salakoski, T. (2004a). Extracting protein-protein interaction sentences by applying rough set data analysis. In Tsumoto, H., Slowinski, R., Komorowski, J., and Grzymala-Busse, J. W., editors, *Proceedings of the Fourth International Conference on Rough Sets and Current Trends in Computing, Uppsala, Sweden*, volume 3066 of *Lecture Notes in Computer Science*, pages 780–785. Springer, Heidelberg.

- Ginter, F., Pahikkala, T., Pyysalo, S., Tsivtsivadze, E., Boberg, J., Järvinen, J., Mylläri, A., and Salakoski, T. (2005a). Information extraction from biomedical text: The BioText project. In Langemets, M. and Priit, P., editors, *Proceedings of the Second Baltic Conference on Human Language Technologies (HLT'05), Tallinn, Estonia*, pages 131–136.

- Ginter, F., Pyysalo, S., Björne, J., Heimonen, J., and Salakoski, T. (2007a). BioInfer relationship annotation manual. Technical Report TR 806, Turku Centre for Computer Science (TUCS).

- Ginter, F., Pyysalo, S., Boberg, J., Järvinen, J., and Salakoski, T. (2004b). Ontology-based feature transformations: A data-driven approach. In Vicedo, J. L., Martínez-Barco, P., Muñoz, R., and Saiz Noeda, M., editors, *Proceedings of the 4th International Conference EsTAL'04, Alicante, Spain*, volume 3230 of *Lecture Notes in Computer Science*, pages 279–290. Springer, Heidelberg.

- Ginter, F., Pyysalo, S., Boberg, J., and Salakoski, T. (2006). Regular approximation of Link Grammar. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors, *Proceedings of the 5th International Conference on Natural Language Processing FinTAL'06, Turku, Finland*, volume 4139 of *Lecture Notes in Artificial Intelligence*, pages 564–575. Springer, Heidelberg.

- Ginter, F., Pyysalo, S., and Salakoski, T. (2005b). Document classification using semantic networks with an adaptive similarity measure. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05), Borovets, Bulgaria*, pages 204–211. Incoma, Bulgaria.

- Ginter, F., Pyysalo, S., and Salakoski, T. (2007b). Document classification using semantic networks with an adaptive similarity measure. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005*. John Benjamins, Amsterdam, The Netherlands.

- Ginter, F., Suominen, H., Pyysalo, S., and Salakoski, T. (2008). Combining hidden markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

- Haverinen, K., Ginter, F., Pyysalo, S., and Salakoski, T. (2008). Accurate conversion of dependency parses: targeting the stanford scheme. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

- Heimonen, J., Pyysalo, S., Ginter, F., and Salakoski, T. (2008). Complex-to-pairwise mapping of biological relationships using a semantic network representation. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

- Laippala, V., Ginter, F., Pyysalo, S., and Salakoski, T. (2008). Resource-efficient construction of a full parser for Finnish nursing narratives. In

Proceedings of the First Louhi Conference on Text and Data Mining of Clinical Documents. To appear.

- Pahikkala, T., Pyysalo, S., Boberg, J., Järvinen, J., and Salakoski, T. (2008). Matrix representations, linear transformations, and kernels for natural language processing. *Machine Learning*. To appear.

- Pahikkala, T., Pyysalo, S., Boberg, J., Mylläri, A., and Salakoski, T. (2005a). Improving the performance of bayesian and support vector classifiers in word sense disambiguation using positional information. In Honkela, T., Könönen, V., Pöllä, M., and Simula, O., editors, *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 90–97, Espoo, Finland. Helsinki University of Technology.

- Pahikkala, T., Pyysalo, S., Ginter, F., Boberg, J., Järvinen, J., and Salakoski, T. (2005b). Kernels incorporating word positional information in natural language disambiguation tasks. In Russell, I. and Markov, Z., editors, *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS'05)*, pages 442–447, Menlo Park, Ca. AAAI Press.

- Pyysalo, S., Ginter, F., Pahikkala, T., Koivula, J., Boberg, J., Järvinen, J., and Salakoski, T. (2004). Analysis of Link Grammar on biomedical dependency corpus targeted at protein-protein interactions. In Collier, N., Ruch, P., and Nazarenko, A., editors, *Proceedings of Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA'04), Geneva, Switzerland*, pages 15–21.

- Pyysalo, S., Sætre, R., Tsujii, J., and Salakoski, T. (2008). Why biomedical relation extraction results are incomparable and what to do about it. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

- Suominen, H., Pyysalo, S., Ginter, F., and Salakoski, T. (2008a). Automated text segmentation and topic labeling of clinical narratives. In *Proceedings of the First Louhi Conference on Text and Data Mining of Clinical Documents*. To appear.

- Suominen, H., Pyysalo, S., Hiissa, M., Ginter, F., Liu, S., Marghescu, D., Pahikkala, T., Back, B., Karsten, H., and Salakoski, T. (2008b). Performance evaluation measures for text mining. In Song, M. and Wu, Y.-F., editors, *Handbook of Research on Text and Web Mining Technologies*. To appear.

- Tsivtsivadze, E., Pahikkala, T., Pyysalo, S., Boberg, J., Mylläri, A., and Salakoski, T. (2005). Regularized least-squares for parse ranking. In Famili, A. F., Kok, J. N., Peña, J. M., Siebes, A., and Feelders, A. J., editors, *Proceedings of the 6th International Symposium on Intelligent Data Analysis*, volume 3646 of *Lecture Notes in Computer Science*, pages 464–474, Berlin. Springer.

Co-edited conference proceedings

- Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T., editors (2006). *Proceedings of the Fifth International Conference on Natural Language Processing FinTAL 06, Turku, Finland*, volume 4139 of *Lecture Notes in Artificial Intelligence*. Springer, Heidelberg.

- Salakoski, T., Rebholz-Schuhmann, D., and Pyysalo, S., editors (2008). *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*. To appear.

Contents

Part I

Research summary

Chapter 1

Introduction

1.1 Background

The shift from the pre-genomic to the post-genomic era, marked chiefly by the sequencing of the human genome (Venter et al., 2001), has been accompanied by a shift in the techniques, aims and challenges of biomedical research. Developments in experimental technology have allowed the large-scale study of not only genomes but also proteins and their interactions, giving rise to the fields of proteomics and interactomics and opening for the first time the possibility of reaching a wide, systemic understanding of organisms. New analysis techniques have also caused a deluge of experimental data, and in response led to broadening collaboration between biologists and computer scientists and increasing research in biomedical data mining and bioinformatics.

Today, it is increasingly recognized that this excess of information extends beyond experimental data also to the biomedical scientific literature in a way that calls for new approaches to dealing with scientific knowledge. The amount of information available in literature collections is already well past the ability of researchers to compose into a coherent whole, and growing at an unprecedented rate. Thus the bottleneck to understanding biological systems is shifting from their analysis to the ability to make use of the results. Despite efforts to gather facts in specialized databases, most information can only be found in scientific publications, fragmented and accessible only to those capable of processing human language.

In response to this information overload there has been an explosion of research in the vibrant young field of biomedical text mining, where methods from machine learning, computational linguistics and computer science are applied to unlock access to the wealth of knowledge generated in the study of genetics, biochemistry, and cellular and molecular biology — knowledge to which human language is both the primary interface and the greatest bar-

rier. It is at this busy intersection of fast-moving sciences that the research described in the present thesis has been done.

The key promise that text mining holds for the biomedical domain is to address the information overload problem by automating the process of "understanding" the relevant parts of the scientific literature. Such automation could, for example, greatly increase the efficiency of searching for information, facilitate the creation of large-scale models of the relationships of biomedical entities, and allow for automated inference of new information as well as hypothesis generation to guide biomedical research.

The following sections describe the challenges, promises and tools of biomedical text mining in more detail and then briefly present one view of a biomedical text mining system before proceeding to describe the publications that form the backbone of this thesis.

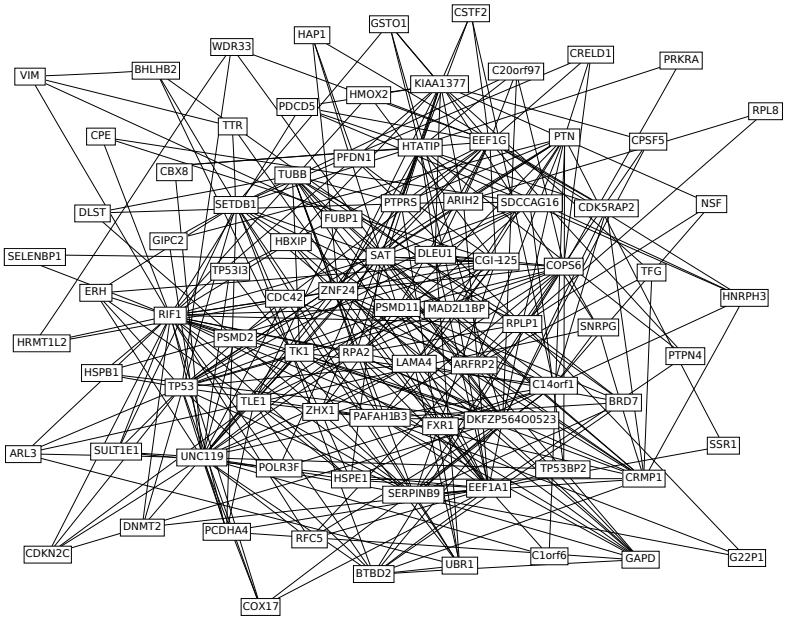

Figure 1.1: *A fragment of the High Confidence interaction network of Stelzl et al. (2005), showing 354 interactions between 90 proteins.*

1.1.1 The need for text mining

Organisms are the most complex systems known, and a full understanding of their function is one of the major goals of present-day biology. This task requires not only an understanding of genes and the proteins they code for, but also of the ways in which they interact. The numbers involved are large and to some extent unknown: for one of the simpler, better-known model organisms, the yeast *Saccharomyces cerevisiae*, the number of genes was famously estimated to be around 6000 upon completion of sequencing (Goffeau et al., 1996)[1] and a recent estimate places the number of protein-protein interactions at 16,000 (Grigoriev, 2003). The numbers for human are more approximate, but one estimate gives the figures 25,000 genes and 375,000 interactions (Ramani et al., 2005). The complexity is vividly illustrated by interaction networks such as the fragment of the human protein-protein interaction network shown in Figure 1.1, illustrating approximately 0.1% of the full human network[2].

[1]The *Saccharomyces* Genome Database currently places the number at 5749; see `http://www.yeastgenome.org/SGD-FAQ.html`

[2]Efforts to map protein-protein interactions are largely focused on determining whether an interaction exists or not, but for a detailed understanding, differences in reactions in response to variation in factors such as cell type and subcellular location must additionally be taken into account, adding another level of complexity to the task.

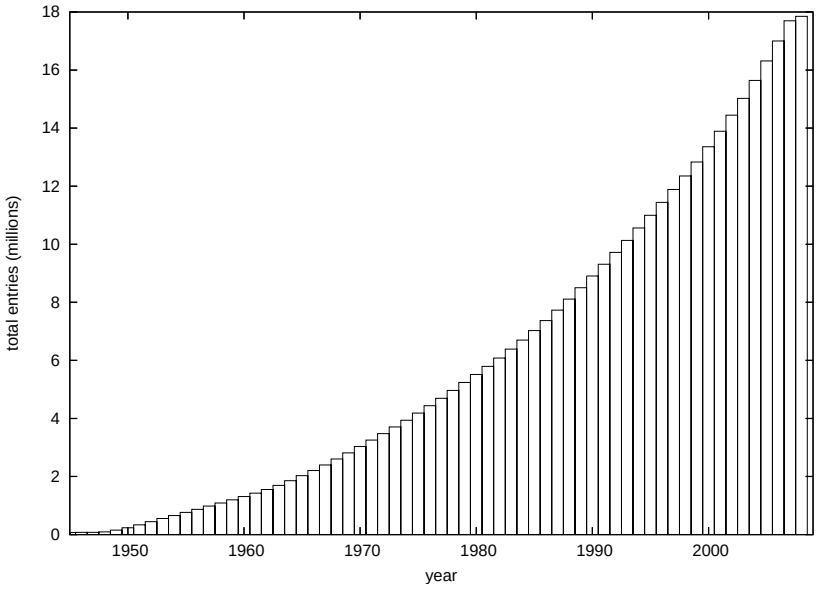

Figure 1.2: *PubMed growth 1945–2008, showing the total number of entries in the database up to the end of each year. Data for 2008 is incomplete.*

The inherent complexity of the objects of study is reflected in the size of the domain literature. The premier literature collection in the biomedical domain is PubMed, which currently contains approximately 17 million citations from five thousand scientific journals, including 9 million article abstracts and, through PubMed Central, 1.5 million full-text articles. For finding relevant information from this formidable collection, PubMed includes an advanced (though ultimately keyword-based) search engine. As an example of both the amount of available information and the insufficiency of naïve keyword search, the name of the protein *p53* occurs in 45,000 PubMed articles, and while a researcher interested specifically in its role in cancer and its interacting partners might try the search *p53 cancer interaction* to narrow down the results, this query still yields 1500 publications, enough for months of full-time reading.

The information overload problem is getting worse, as the biomedical research literature is growing at a daunting rate: almost 700,000 references were added to PubMed during 2007, for an average of 1900 per day. The growth of PubMed, illustrated in Figure 1.2, is double-exponential: even the growth rate is increasing exponentially, at an estimated 3% compounded annual rate (Hunter and Cohen, 2006). The ongoing Open Access revolution in scientific publishing (see e.g. Suber, 2002) is further extending routine free access from article abstracts to full text articles, considerably increasing the growth rate of available textual information.

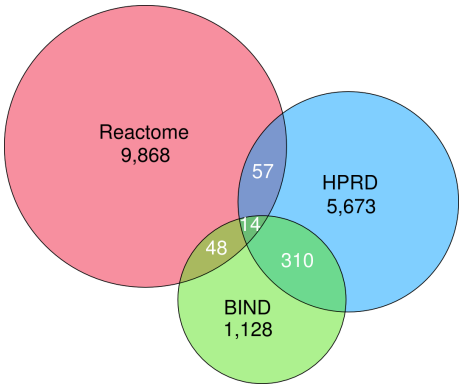

Figure 1.3: *Overlap between existing human protein interaction sets in the Reactome (Joshi-Tope et al., 2005), Human Protein Reference Database (HPRD; Peri et al., 2004) and Biomolecular Interaction Network Database (BIND, Bader et al., 2001) datasets. The small overlap (< 0.1% in common in all three datasets) implies that the number of protein interactions described in the literature is actually quite large and that the individual datasets carry specific biases. (Figure and caption adapted from Ramani et al., 2005, reproduced under the Creative Commons Attribution License)*

In response to this complexity, a number of projects have been initiated to collect information from the diverse publications into special-purpose databases for researchers. These efforts have supplemented well-established resources focusing on information regarding individual biomedical entities (e.g. SWISS-PROT; Boeckmann et al., 2003) by creating resources focused on their relationships, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa and Goto, 2000), the Database of Interacting Proteins (DIP, Xenarios et al., 2000), the Biomolecular Interaction Network Database (BIND, Bader et al., 2001), the Molecular INTeraction database (MINT, Zanzoni et al., 2002), and EcoCyc (Karp et al., 2002). These manually curated databases today contain references to tens of thousands of protein-protein interactions described in the literature and have become important resources for investigating biological systems.

However, manually curated databases still only cover a small fraction of published interactions. Recent studies by Ramani et al. (2005) and Mathivanan et al. (2006) observed this through the small overlap between the interaction annotations in different databases (see Figure 1.3). Further, it is not clear whether the gap between the interactions reported in the literature and found in databases is narrowing or growing. In a recent study on genomic knowledge base construction, Baumgartner et al. (2007) argue that the present rate of manual curation is insufficient to functionally annotate even currently available proteomes.

1.1.2 The promise of text mining

Text mining promises to relieve the information overload problems by allowing facts to be automatically extracted from text. The use of automated methods to assist in uncovering facts that are stated in a body of biomedical literature too large to be practically analysed by humans predates the current wave of domain studies: the pioneering Arrowsmith system of Swanson (1986), based on the co-occurrence of medical domain terms in article titles, predicted an association between fish oil and Raynaud's disease as well as a link between magnesium deficiency and migraine headaches (Swanson, 1988); both of these hypotheses were later verified experimentally.

Despite its successes, the approach of Arrowsmith is based on simple word statistics, which have limited value in analyzing the meaning of text. In the intervening two decades, considerable advances have been made in the field of Natural Language Processing (NLP), allowing much more sophisticated approaches to be brought to bear on text mining today. In particular, techniques developed for a form of *natural language understanding* offer the possibility of automating many of the tasks involved in creating large-scale biomedical databases of facts from the domain literature. Natural language understanding is a long-standing, long-term goal of much of NLP research. The aim is to create systems that take unrestricted natural language input and analyse its entire meaning, capturing it in a representation that can further support inference of facts entailed but not directly stated by the input. This ambitious goal is widely considered AI-complete[3] and not likely to be solved in the near future. For practical applications, meeting the more modest and realistic goal of restricted understanding of particular types of statements in a given domain is already valuable. This type of limited natural language understanding, producing a structured but not necessarily computation-supporting representation of facts, is referred to as *Information Extraction* (IE).

Much of the recent work in biomedical NLP pursues goals that can be viewed either as full IE problems or subtasks of an IE problem. Domain extraction targets include, among others, relations between genes and drugs (Rindflesch et al., 2000b), genes and mutations (Rebholz-Schuhmann et al., 2004), treatments and diseases (Rosario and Hearst, 2004), genes and diseases (Chun et al., 2006), and, for example, statements regarding protein localization (Craven and Kumlien, 1999) and protein active sites (Gaizauskas et al., 2003). While binary relations are by far the most common target, some methods targeting n-ary relations describing e.g. gene variation (Mc-

[3]The informal term *Artificial Intelligence-complete* refers to the class of problems that would require human-level intelligence to truly solve — and *"if we could solve any one artificial intelligence problem, we could solve all the others"* (Mueller, 1987, page 302); the analogy is with the computational complexity class of NP-complete problems.

Donald et al., 2005) and subcellular localizations (Melli et al., 2007) have also been presented. The most common goal for domain IE research is the recognition of protein-protein interactions (PPIs), and the most commonly addressed subtask is the recognition of protein names[4]. PPIs are also the primary extraction goal of the studies described in this thesis.

Efficient, scalable, high-reliability PPI extraction would clearly be a boon for efforts to build protein-protein interaction networks and annotate protein databases with information regarding protein functions. Such systems could also be used to augment literature searches, assist in the analysis of the results of experiments, and serve as building blocks for hypothesis generation systems. While the state of the art has yet to reach a point where these promises could be fully realized, many systems have already been deployed and several have been shown to provide practical benefits for database curators (see e.g. Donaldson et al., 2003; Müller et al., 2004; Couto et al., 2006; Ohta et al., 2006; Alex et al., 2008; Karamanis et al., 2008; Kim et al., 2008b). The following section examines some of the challenges in biomedical text mining.

1.1.3 Challenges in text mining

The PPI extraction task is typically cast as a problem of finding pairs of proteins that are stated to interact in a given text. At first sight, this may appear deceptively easy: the minimum requirements are recognizing protein names and making a decision for each pair whether or not they interact. Several resources containing protein names are available (e.g. Bairoch et al., 2005; Liu et al., 2006), and interactions are often stated through verbs such as *bind* and *phosphorylate*, which can further be recognized with statistical methods (Andrade and Valencia, 1998). The problem would thus appear to reduce to dictionary lookup combined with some form of pattern matching.[5]

A number of early PPI extraction methods followed this approach, employing fixed lists of protein names and making use of sentence structure only to the extent of looking at the order of words and the distance between them to match simple patterns such as "protein1 — action — protein2" (see e.g. Blaschke et al., 1999). While some early results using this approach on narrow subdomains and restricted classes of interaction types were very encouraging, results from larger, more realistic settings have confirmed that simple pattern-based approaches do not achieve sufficient performance.

[4]The recognition of names can be viewed an IE task in its own right. As is common in domain studies, the term will here be used to refer to what McNaught and Black (2006) (page 149) term *"higher"* IE tasks, i.e. those involving relations between entities.

[5]The approach pursued in most PPI extraction work, including that in this thesis, relies in some way on sentence structure. Purely statistical approaches such as that of Arrowsmith have also been applied in the biomedical domain (see e.g. Stapley and Benoit, 2000; Jenssen et al., 2001) but will not be considered in more detail here.

Two of the main obstacles in the way of fully automatic extraction of facts from free-form natural language text are *ambiguity* and *variability*. Ambiguity refers to a single expression having multiple interpretations and variability to a single "interpretation" (semantic representation of facts) being denoted by multiple expressions. The two are thus, in one sense, opposites, yet equal in that both greatly complicate the extraction of facts from text.

In everyday communication, language users rarely realize that essentially every nontrivial sentence is ambiguous, that is, has two or more possible interpretations. Indeed, unexpected ambiguity is a frequent source of humor:

> One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know. *(Groucho Marx in the movie Animal Crackers)*

The joke arises from the two interpretations of *shot an elephant in my pajamas*, which exhibits prepositional phrase attachment ambiguity. Ambiguity occurs at multiple levels in language, from the meaning of individual words to syntax to interpretation of syntactic structures: among other issues, complex sentence structure and words that are unknown to text analysis tools, both common features in biomedical text, are rich sources of ambiguity. Compounding the problem, ambiguity is typically combinatorial: two instances of two independent alternatives make for a sentence with four readings, four for 16, etc. The number of alternative analyses thus typically grows exponentially with the number of tokens in the sentence, leading in some cases to enormous numbers of ambiguous alternative interpretations.

Like ambiguity, variability is an essential property of human language. While authors strive for variability and readers abhor repetitive or formulaic text, the creative potential of language and the use of that potential by authors considerably complicates the automatic extraction of facts stated in text. Even a simple relationship such as one protein binding another can be expressed in a surprising number of ways: simplified variants found from a small sample of sentences in the BioInfer corpus (see Paper IV and Chapter 5) are shown in Table 1.1. Each of these statements entails, in context, an actual or possible binding relationship between two biomedical entities. Due to the variability and productivity of natural language, it is not possible to enumerate the complete set of full word sequences that can express interesting facts: most of the original forms from which the examples in Table 1.1 are taken occur only once in the corpus, and the flexibility of natural language guarantees that novel variants will occur, escaping any collection of known forms: the distribution of patterns, like many other features in natural language, exhibits characteristics that can be described in terms of Zipf's law (see e.g. Rebholz-Schuhmann et al., 2005).

e_1 binds e_2	e_1 cross-links e_2
binding of e_1 to e_2	e_1 binding to e_2
binding to e_1 by e_2	e_1 is able to bind to e_2
e_1 is shown to bind e_2	e_1 is an antigen known to bind e_2
e_1 (an e_2-binding protein)	partners that associate with e_1: e_2 and e_3
e_1 is involved in binding to e_2	e_1 has been implicated in e_2 binding
e_1 binding region of e_2	e_1 is secreted as a protein that binds e_2
affinity of e_1 for e_2	e_1 regulates e_3 by binding to e_2
association of e_1 with e_2	e_1 is directly associated with e_2
e_1, which binds to e_2,	e_1 is a receptor for e_2
e_1 binding sites of e_2	e_1 is expressed as a receptor for e_2

Table 1.1: Twenty-two ways to say e_1 *binds* e_2.

Thus, a collection of exact word sequence patterns for PPI extraction will suffer from insufficient coverage and hence poor recall. One response is to allow flexibility: for example, extract an interaction whenever a word commonly used to express interactions occurs within a given distance from two protein names (Blaschke et al., 1999). However, such approaches fail in many cases: consider for example the sentence

> By site-directed mutagenesis of e_1 from *Dictyostelium discoideum* the point mutations K114E and W3N were generated by PCR, thus changing e_2 and e_3-binding activity, respectively.

Here, the pairs (e_1, e_2) and (e_1, e_3) are stated to bind, but (e_2, e_3) is not. A keyword position-based approach flexible enough to find the first two would also match the third, thus having limited precision.

A more sophisticated approach is needed to make high-precision, high-recall text mining possible. The approach taken in most text mining systems is to divide the task into a sequence of steps, each performed by a dedicated module devoted to a well-defined subtask. From the perspective of the overall goal of IE, such organization can be viewed as sequential disambiguation and normalization of the input to resolve the ambiguity and variability of the natural language text, thus making the relevant information more readily extractable in a final step. The following sections sketch such a system and relate the research described in this thesis to the relevant stages of processing.

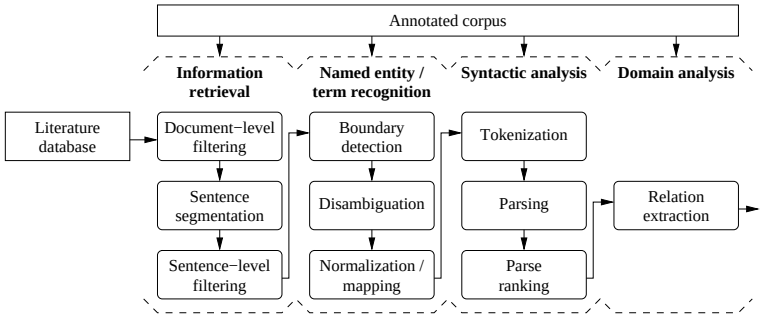

Figure 1.4: *Text mining system architecture.*

1.2 Text mining tasks

This section describes the processing steps of a text mining system as conceived in the projects in which the present work has been carried out. The system described here is conceptual—that is, it has not been realized as a coherent, whole software system, and although several working components exist, the interest in their development is mainly as research prototypes rather than as software development projects. Nevertheless, the design of the system provides a framework into which the discussed modules fit, and the processing stages as well as their associated challenges and current solutions will next be briefly presented to set the stage. Where applicable, the performance[6] of methods on biomedical domain text will be related to performance on "general English" (roughly, newspaper-type text) to highlight specific challenges. Additionally, references to co-authored and other relevant publications produced in the projects in which the present research was carried out are provided to relate these to the papers included in this thesis.

The architecture of the considered text mining system is presented in Figure 1.4. In brief, documents from a literature database such as PubMed are filtered to a set of likely relevant sentences which are then marked for named entities and parsed prior to the extraction of relevant information; the process is supported by an annotated domain corpus. The steps shown in the figure are not exhaustive nor all necessary: commonly applied processing stages not shown include part-of-speech tagging and coreference resolution, and only named entity detection and relation extraction are strictly mandatory. Further, there is no requirement that the steps discussed here be performed in the particular order shown in Figure 1.4. For example, the

[6]Unless noted otherwise, performance is given throughout this thesis using the standard metrics of precision, recall, and balanced F-measure ("effectiveness"; van Rijsbergen, 1979), all of which range from zero to one, the larger the better. *Error* is one minus the contextually relevant metric.

Alvis system (Deriviere et al., 2006) places named entity recognition before sentence segmentation to e.g. avoid splitting sentences on name-internal periods, and Finkel et al. (2004) report that full parsing is beneficial for named entity disambiguation. The order presented below (mainly following Figure 1.4), progressing first from documents to sentences and then from the word level to syntax to (shallow) semantics, is commonly applied.

1.2.1 Information retrieval

Information Retrieval (IR) refers to the task of selecting a subset of relevant items from a large body of data (here text), typically in response to an *ad hoc* query given by a user. This means in effect that the system cannot anticipate the relevance criterion. IR is an extremely well-studied task in which significant practical advantages have been achieved with advances both in the computer science and natural language processing facets of the task—Google probably representing the most widely known example of breakthrough IR technology and PubMed search being the example most relevant to the day-to-day work of biomedical domain researchers.

From the perspective of a PPI extraction system, IR acts in the role of a filter reducing the amount of input and increasing the fraction of relevant information (Hersh et al., 2004). The IR task can then be approached either as a traditional *ad hoc* retrieval problem or as a classification problem with a fixed definition of relevance: for example, all documents describing PPIs are relevant, while others are not. The latter approach allows the problem to be addressed using supervised classification methods, which can achieve considerably better results through the use of labeled training data.

The IR community has a long-standing series of competitive evaluations, the Text REtrieval Conferences (TREC), which have hosted shared IR tasks annually since 1992. A large number of different problem domains, termed tracks, are hosted each year. TREC included a Genomics track which ran from 2003 to 2007. In addition to *ad hoc* retrieval tasks (2003–2005), the Genomics track included tasks on document classification (2004–2005), question answering (2006–2007), and, in 2003, an IE-type task, extraction of GeneRIF (Reference Into Function) text. It should be noted that while this last task included the extraction of text describing interactions, it did not require extraction of structured information but instead of text snippets. Thus, this task should be viewed as a highly focused IR problem rather than as IE in the sense considered here.

During the course of the projects in which this work was carried out, a number of studies addressing IR-type problems were published: a supervised learning approach to distinguish between sentences describing PPIs and those that do not (Pyysalo, 2003), and two studies on the use of ontologies as the basis of document similarity measures that introduced a pair

of novel methods for tuning the measures in a task-dependent way (Ginter et al., 2004b, 2005; Ginter, 2007). Nevertheless, like most studies focusing on PPI extraction, the work described in this thesis mostly assumes that the input to the system consists mainly of relevant documents, that is, that moderately high-precision filtering of the input documents has been performed as a preprocessing step. The significance of this assumption for measured PPI extraction system performance is discussed in Section 6.3 of this thesis.

1.2.2 Sentence segmentation

Sentences are both the most natural and most common choice for the unit of text from which to extract information, reflecting in part the fact that parsers operate on the sentence level. Using sentences as the unit for extraction finds support in the evaluation of Ding et al. (2002), although their estimate of sentence-level recall suggests that when extraction is done only within sentences, approximately 15% of relationships are overlooked unless references between sentences are recovered with a coreference resolution method.

Running text must be *segmented* (split) into sentences prior to parsing and other sentence-level processing. Although both under- and over-segmentation will affect performance adversely, the segmentation task is not widely studied, perhaps due to its relative simplicity. For example, the widely applied MXTERMINATOR segmenter of Reynar and Ratnaparkhi (1997), trained on general English, is used also in a number of biomedical text processing pipelines (e.g. Rinaldi et al., 2006), although in biomedical English periods and capital letters are less reliable indicators of sentence boundaries than in general language: consider e.g. decimal points, abbreviated genera in species' names (*S. cerevisiae*) and abbreviated names of authors and journals in citations.

Two recent studies by Tomanek et al. (2007) and Xuan et al. (2007) focus on the challenges of sentence segmentation in the biomedical domain. Tomanek et al. take a machine-learning based approach, while Xuan et al. develop a rule-based system that makes intensive use of specially collected dictionaries. Both studies report error rates below 0.3% for their proposed methods, outperforming the best reported result for MXTERMINATOR on general English (Reynar and Ratnaparkhi, 1997) by a fair margin. These results appear to restore the status of segmentation as a "solved" problem also for biomedical text. Segmentation has not been specifically considered in any of the studies in this thesis, and perfect segmentation of the input is assumed for the conceptual system in all of the studies.

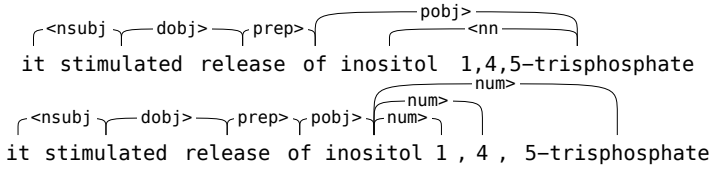

Figure 1.5: *Illustration of the effect of tokenization on parsing. Above: correct tokenization with correct parse. Below: over-split tokens leading to incorrect recognition of head and treatment of token fragments as numerical postmodifiers.*

1.2.3 Tokenization

Tokenization, or *word segmentation*, refers to the splitting of continuous text into tokens, atomic units of text: words, numbers, punctuation etc. Tokenization is another under-appreciated task, often relegated to the generic algorithms included in taggers or parsers. In one of the rare studies focusing on the issue, Tomanek et al. (2007) perceptively describe the prevailing attitude toward segmentation as considering the tasks *"unsophisticated clerical work."* There are, however, specific domain tokenization requirements. Biomedical text includes many syntactically atomic units that contain characters commonly appearing, and assigned to, separate tokens in general text: token-internal periods (*H2A.2*), commas (*4,5-bisphosphate*), parentheses (*poly-(L)-proline*) and slashes (*Arp2/3*), for example, frequently occur as parts of names. By contrast, many units that are commonly preserved as single tokens contain relevant internal structure for semantic interpretation, the most common example being adjectives such as *CREB-binding*.

Tokenization errors are an effective way to confuse parsers, and e.g. the influential Penn tokenization algorithm splits tokens in many of the examples above (see also Figure 1.5). Tomanek et al. study domain tokenization in the light of the annotations of the GENIA Treebank (GTB) (Tateisi et al., 2005) and PennBioIE (Kulick et al., 2004) corpora, and compare their value as training material for a machine learning method for tokenization targeting a semantically motivated tokenization. They find disastrously poor performance, 72% accuracy, when training on the GTB corpus, but achieve almost 97% accuracy when performing cross-validation on the target corpus.

However, the fact that the GTB tokenization recognizes syntactically, not semantically, motivated tokens is not necessarily cause for criticism—GTB is, after all, a treebank. Tokenization for parsing purposes is not a particularly difficult task: even the Penn tokenization script[7], not in any way optimized for biomedical text, creates 98.5% of the same tokens on the BioInfer corpus as in our manually corrected, syntax-oriented tokenization; minor tweaks suffice to bring this figure past 99%. There is no need to

[7]Available from `http://www.cis.upenn.edu/~treebank/tokenization.html`

treat, for example, *Arp2/3-binding region* as more than two tokens in order to parse it. Identifying the units that are necessary for correctly determining the meaning (minimally, *Arp 2 / 3 -binding region* to resolve *Arp2*, *Arp3*, *binding* and *region*) is evidently a harder problem, but creating these tokens early in the processing pipeline could unnecessarily complicate parsing. One alternative is to split the "atoms" of syntax further for semantic processing: other than possible technical restrictions, there is no requirement that the syntactic token match the semantic "token," similarly as morphemes, the elements forming words, are distinct from the elements of syntax (see e.g. Mel'čuk, 1988, chap. 3). Grover et al. (2005) discuss multiple levels of tokenization specifically in the context of parsing biomedical text. We have followed this approach in our recent work, splitting e.g. *actin-binding* into the tokens *actin* - *binding* and introducing additional "dependencies" between them after parsing (Airola et al., 2008).

While tokenization for parsing is adequately addressed by simple rules, the indication that machine learning for semantically motivated tokenization has an over 3% error rate suggests that this specific problem may deserve more consideration. As the issue is essentially word-internal, the standard tools of computational morphology—one of the solved problems in natural language processing (Karttunen, 2007)—should provide a satisfactory solution with some investment of manual work. Tokenization and the multiple levels of tokens in the BioInfer corpus are discussed further in Section 5.1.

1.2.4 Word-level processing

To normalize words, it is common in many applications to apply a morphological analyser to retrieve the lemmas (base forms) of words with tags denoting their parts-of-speech (POS) and inflectional and other features. However, due perhaps to the extreme poverty of syntactically-driven morphology in English (compared to e.g. Finnish), many domain studies either do entirely without lemmatisation or use simple general English stemmers such as that of Porter (1980). The subtask of POS tagging, however, has been studied in the domain. POS tagging involves assigning to each word the most likely tag (or tags) from a predefined set denoting parts of speech (determiner, noun, adjective etc.). Numerous well-developed machine learning techniques for tagging exist, and POS annotation is not as demanding to create manually as e.g. full syntactic analysis or semantic annotation. Following the introduction of POS-annotated domain corpora, several high-quality biomedical domain POS taggers have been introduced: The MedPost tagger of Smith et al. (2004) and the GENIA tagger of Tsuruoka et al. (2005) both achieve over 97% accuracy on biomedical text, essentially matching the performance of machine learning-based POS taggers on general domain English. (see e.g. Shen et al., 2007).

Some caveats apply to this simplified view of POS tagging performance. First, tagging performance in excess of 99% has been reported for the hand-written rule-based EngCG system (Voutilainen et al., 1992; Voutilainen, 1995) in a relatively early study by Samuelsson and Voutilainen (1997). While differences in tagset restrict the ability to make a direct comparison and the degree to which this level of performance generalizes to other domains has been questioned (Entwisle and Powers, 1998), accuracy of nearly 98% on the GENIA corpus has been reported for a rule-based system (Castaño and Pustejovsky, 2005), suggesting that the level of domain tagging performance could be further improved with a custom hand-written system. Nevertheless, as the performance of automatically derived taggers is relatively high and there is considerable effort involved in the creation of a system such as EngCG, this may not represent a good investment of development effort. Second, as the common word-sequence approach to POS tagging gives a lower-level view of sentence structure than that available to parsers, many state-of-the-art parsers do not perform POS tagging prior to parsing but rather incorporate tagging into the parsing process. The "standalone task" view of tagging performance above does not consider full parsing, as is appropriate for e.g. tasks that benefit from POS tagging but for which full parsing may be computationally too expensive.

POS tagging does not appear to present a bottleneck for text mining performance and has not been considered as a primary goal in the research presented in this thesis. However, domain taggers are considered as a foundation to build on, and the value of tagging in parser domain adaptation is studied in Paper II and discussed in more detail in Chapter 3 of this thesis.

1.2.5 Named entity recognition

Recognizing the names of entities in text is a fundamental prerequisite for the extraction of information regarding their relationships. The task can be divided into three subtasks: *named entity detection*, where the occurrences of names are marked, *named entity classification* (or disambiguation), where the type of the named entity (e.g. gene or protein) is determined, and *named entity normalization*, where spelling variants are normalized to determine the canonical name of the named entity and, typically, relate it to an entry in a database such as UniProt (Bairoch et al., 2005). These steps fall broadly under the heading of *Named Entity Recognition* (NER), although not all NER methods attempt all these subtasks. In the following, the common practice of using NER as a catch-all term for tasks where at least detection is performed will be followed, with specific subtasks identified when relevant.

Biomedical domain NER is particularly challenging for a number of widely recognized reasons. Ambiguity is pervasive: in addition to gene names and synonyms that are common English words (notoriously common

Ambiguity with general closed-class English words:				
an	by	can	for	not
Ambiguity with other domain words:				
head	blood	cell	double	arm
spliced	eyeless	limited		
Ambiguity with out-of-domain English:				
kayak	canoe	midget	rutabaga	18-wheeler
vamp	ogre	disco	boss	shaggy
Spelling variation:				
RAR alpha	RAR-alpha	RARA	RARa	RA receptor alpha
NF-kappaB	NK(kappa)B	kappaB	NF-KB	NFKB factor

Table 1.2: Examples of problematic gene and protein names and abbreviations. (Examples from Proux et al., 1998; Hirschman et al., 2002a; Leser and Hakenberg, 2005; Yeh et al., 2005; Ananiadou and Nenadic, 2006)

in particular in the *Drosophila* nomenclature) there is systematic ambiguity between gene and protein names that arises from the close relatedness of genes with the proteins they codes for; often the same name is used for both. Other issues include spelling variation, inconsistent use of abbreviations, nested names, synonyms (arising from e.g. independent discovery), descriptive names, and the constant discovery and naming of new genes. Table 1.2 illustrates a number of cases; Park and Kim (2006) and Ananiadou and Nenadic (2006) discuss these issues in detail.

Due to its importance to all further processing, NER is one of the most widely studied tasks in biomedical text processing. While some initial results for rule-based systems were very promising (Fukuda et al., 1998; Proux et al., 1998) and a comparative evaluation by Nobata et al. (2000) suggested the problem would not be notably more difficult for supervised learning methods than general English NER, broader studies have revealed the task to be very challenging. As NER has been a target in a number of shared task evaluations, state-of-the-art performance will here be approached through the results in the JNLPBA (Kim et al., 2004), BioCreative (Hirschman et al., 2005b) and BioCreative II (Wilbur et al., 2007) evaluations.

The JNLPBA shared task used a version of the GENIA corpus (Ohta et al., 2002), simplified to remove nested types and restricted to five of the 36 annotated classes: protein, DNA, RNA, cell line and cell type. The task thus required detection and disambiguation. The BioCreative evaluation task 1A (gene mention finding) was a single-class named entity detection problem (Yeh et al., 2005) using the GENETAG corpus, tagged for gene and protein names including related domains, complexes, subunits and promoters (Tanabe et al., 2005). The BioCreative II GM corpus was an extension of

	Best result			
Evaluation	precision	recall	F-measure	
BioCreative, 1A (closed)	82.0%	83.2%	82.6%	(Zhou et al., 2005)
BioCreative, 1A (open)	82.8%	83.5%	83.2%	(Finkel et al., 2005)
JNLPBA	69.4%	76.0%	72.6%	(Zhou and Su, 2004)
BioCreative II, GM	88.5%	86.0%	87.2%	(Ando, 2007)

Table 1.3: Best named entity detection shared task results. BioCreative submissions were categorized as either *closed* or *open* depending on whether they applied only the given training data or also other resources such as lists of gene names; no restrictions were placed on BioCreative II submissions. The JNLPBA task required also disambiguation between five entity types.

the BioCreative corpus, with 50% more training data. Table 1.3 shows the best results from the shared task evaluations. One interesting observation from these results is that the use of external resources such as gazetteers (dictionaries of names—critical resources in general English NER) brings relatively little extra value to the biomedical task. The surprising finding of lower performance for the JNLPBA evaluation, which followed BioCreative, is explained in part by the requirement for disambiguation, the inclusion of partial matching in BioCreative 1A (Tanabe et al., 2005) and possibly also in part by annotation consistency (Dingare et al., 2005). Tsai et al. (2006) discuss these issues in detail.

The improvement between the two BioCreative evaluations is encouraging (though a number of confounds such as training set size exist), and the organizers of the latter challenge were further able to demonstrate that a combination of all system outputs, while not a practically workable approach, could in theory achieve 90.7% F. Nevertheless, error rates remain considerably larger than the best achieved for general English: in the sixth Message Understanding Conference, the first to introduce NER as a distinct subtask, the highest-performing system achieved 96% recall and 97% precision, matching human performance (i.e. interannotator agreement rate) (Sundheim, 1995). Tests where the same method is applied to general and biomedical domain data (e.g. Finkel et al., 2005) confirm that the biomedical problem is indeed more difficult, and interannotator agreement results (e.g. Gaizauskas et al., 2003) suggest that this holds also for human annotators (part of this difference may be related to the longer average length of biomedical names; see Yeh et al., 2005).

The task of disambiguating between different classes of biomedical names can either be performed as part of named entity detection (e.g. as multiclass B-I-O tagging; Kim et al., 2004) or as a separate step. The latter approach was studied early on by Hatzivassiloglou et al. (2001), who reported an F-measure of 85% for disambiguating between gene and protein names. We

have considered gene/protein disambiguation as a model task in a number of studies exploring context representation models for machine learning for disambiguation-type tasks. The original intuition in these studies was that the common strategy of extracting features for learning from a limited window of context words represents an unrealistic model of context relevance, where the closest words (those within the window) are all equally relevant and others completely irrelevant. This model has been used in a great number of NER studies, including specifically in biomedical NE disambiguation experiments by Hatzivassiloglou et al. (2001) and Torii et al. (2003, 2004). This hard-boundary model of relevance was replaced with a distance-based decay function that was initially shown to outperform other methods at this task when applied together with a custom classifier by Ginter et al. (2004a). In follow-up studies, context weighting approaches were shown to be beneficial also for gene/protein disambiguation with Support Vector Machines (SVM) (Pahikkala et al., 2005a), and general word sense disambiguation tasks with both SVM and Naïve Bayes classifiers (Pahikkala et al., 2005b). We have further generalized context weighting and the incorporation of positional information of context words in two studies that develop the idea toward a framework of kernels for disambiguation tasks (Pahikkala et al., 2005c, 2008). These general methods are potentially applicable also to numerous other tasks, including named entity detection using the B-I-O model.

The final NER subtask, normalization, is also referred to as *grounding* or *mapping* when understood to involve the identification of a unique identifier corresponding to the named entity in a biomedical domain database. The normalization problem is somewhat specific to biology, as general English names feature only a fraction of the variability seen in e.g. protein names. Grishman (2003) does not consider the subtask in his treatment of NER for IE, and normalization was not included in the influential Message Understanding Conference (MUC) series formulation of NER, although alias recognition in MUC-6 template element filling (see e.g. Grishman and Sundheim, 1996) can be seen as a related problem and the *"off the page"* (Doddington et al., 2004) aims of the later ACE Entity Detection and Tracking tasks pose somewhat similar challenges. The BioCreative evaluation task 1 included a normalization subtask (Hirschman et al., 2005a), where the best achieved results were 92% F-measure for normalizing yeast genes, 82% for fly, and 79% for mouse. BioCreative II only tested human genes, with the best system achieving 81% and the organizers 84% by pooling system outputs (Morgan and Hirschman, 2007). These results demonstrate strong subdomain dependence for the task, and this variation further complicates comparison to performance in other domains. While the BioCreative II normalization task organizers judged the results to indicate *"a significant advance in the state of the art"* (comparing with previous mouse results and taking the

higher ambiguity of human gene names into account), the normalization task, which requires named entity detection, holds considerable challenges. Normalization is not studied as a specific goal in this thesis, and named entity identity is intentionally excluded from consideration in particular in relation extraction tasks.

1.2.6 Term recognition

In the biomedical domain, the task of *term recognition* is in many ways closely related to NER; indeed, term recognition can be viewed as encompassing named entity recognition (see e.g. Krauthammer and Nenadic, 2004; Ananiadou and Nenadic, 2006) and how the borders are drawn may depend largely on methodology, with IE and machine learning providing one view and automatic term recognition and a linguistic approach another. The boundaries of names and, consequently, of the named entity recognition task are relatively crisp in general English (Sundheim and Chinchor, 1995). The core of NER is the task of tagging proper names, that is, unique (in context) labels of either specific individuals or groups such as families, as opposed to classes of entities. Proper names do not, as a rule, describe their referents (for example, *Philip Kindred Dick* does not imply fondness of horses, relatedness, or fatness) and they are not compositional: the meaning of a name is not related to the meanings of its parts. Biomedical names do not share these properties: protein names almost never identify an individual protein, but rather a class of entities, names often derive from descriptions (e.g. *CREB-binding protein*) and frequently follow compositionality: for example, the protein *MAP* is phosphorylated by *MAP kinase* (kinase meaning a phosphorylating enzyme), which is in turn phosphorylated by *MAP kinase kinase*.

The view we have taken especially in Paper IV (see also Ginter et al., 2007) is a pragmatic one: protein and gene names are established labels of (mutually interchangeable classes of) biomedical entities as used by biologists and represented in domain resources such as UniProt. In this view the possible internal structure of names is a coincidental, not fundamental property, and the main goal of normalization is to establish identity. By contrast, the internal structure of terms and the way in which it represents their relations (e.g. of *actin filaments* and *filaments*) is central to terminological processing, and the aim of term mapping is to relate terms to an ontology of concepts that defines relationships such as meronymy (part-whole) and hyponymy (class-subclass).

While term recognition, distinguished from NER in this way, is not specifically considered as a target by any of the methods introduced in the studies discussed in this thesis, terms are relevant to the annotation of the BioInfer corpus (Paper V), discussed in Chapter 5.

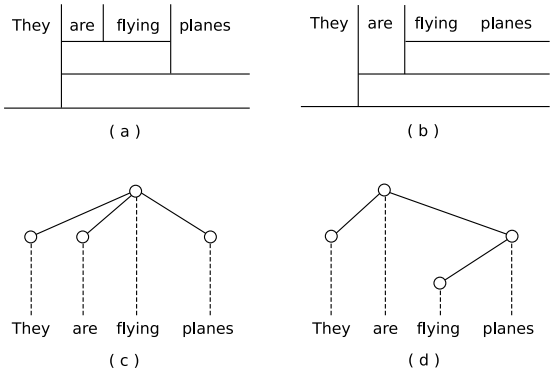

Figure 1.6: *An early view of a pair of constituency (a,b) and dependency (c,d) analyses for an ambiguous sentence. (Figure adapted from Hays, 1964)*

1.2.7 Parsing

Syntactic analysis, or parsing, refers to the task of analyzing sentence structure to produce a representation of its syntax. Parsing can be divided into two broad traditions: *constituency* (or *phrase structure*) and *dependency*. The first holds that words combine into phrases which repeatedly combine to form the sentence. The motivation for constituency analysis arises from the observation that there are combinations of words, phrases, that can be substituted for single words: for example, the noun *proteins* can be substituted with the noun phrase *50-kDa proteins derived from Acanthamoeba*. By contrast, the dependency view rejects intermediate levels of description and instead analyses syntax as binary relations, dependencies, that hold between words. The motivating observation is that words depend on other words, such as a determiner on a noun, for their presence in the sentence.

Modern constituency-based syntactic theories can be traced back at least to the formulation of Bloomfield (1933) and for computational, generative approaches to the enormously influential work of Chomsky (1957). A dependency view of sentence structure can be traced back centuries, and much of modern dependency theory to the seminal work of Tesnière (1959) and the mathematical formulations provided by Hays (1964) and Gaifman (1965); an excellent recent advocacy of dependency syntax is given by Mel'čuk (1988). An early view of constituency and dependency analyses is illustrated in Figure 1.6. In present-day computational linguistics, the most influential constituency scheme is that of the Penn Treebank (PTB) (Marcus et al., 1993), which has been almost universally adopted in particular in statistical parsing studies. While no equivalent standard exists for dependency, the Stanford dependency scheme (de Marneffe et al., 2006) has many typical characteristics. A PTB tree and a Stanford dependency analysis of a sentence are shown in Figure 1.7.

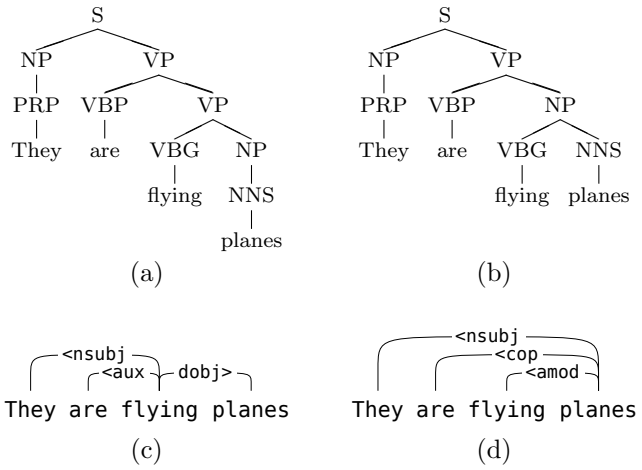

Figure 1.7: *Example Penn Treebank (a,b) and Stanford dependency (c,d) analyses.*

Constituency has long been dominant in mainstream theoretical and computational linguistics, but there has recently been a resurgence of approaches based on dependency, driven in part by increasing interest in languages other than English and focus on semantics. The focus of the work presented in this thesis is on dependency. The parsers considered here produce full (as opposed to partial) dependency representations of sentence structure, with varying emphasis on surface syntactic versus deep structure.

In dependency theory, certain properties are commonly understood to constrain valid syntactic structures. Most fundamental among these are that no word depends on more than one other word, called its *head* (single-headedness), and that each word except one, the root, depends on another word. The relation is represented by a directed edge from the head to the dependent, and the structure is a rooted tree. Each head is said to *govern* its *dependent*. Some formalizations add constraints such as *projectivity* (Hays, 1964; Gaifman, 1965): informally, that the dependencies must not cross when drawn connecting the words of the sentence. However, projective dependency grammars are less powerful than those without this constraint, and projectivity has been argued to restrict the ability of the formalism to describe natural language (see e.g. Covington, 1990; Tapanainen and Järvinen, 1997).

Common usage, in particular in computational linguistics, understands the concept of dependency broadly. In this thesis, an inclusive view is adopted: dependencies are, roughly, understood to be binary relations that hold between words. Dependencies are typically typed and are either explicitly directed or provide some other means (e.g. type and word order) for determining the roles of the words connected by the dependency. This broad

view allows for multiple heads to model e.g. control phenomena without explicitly separated dependency layers (strata) and includes Link Grammar, among others, in the class of "dependency-type" formalisms.

There are connections between the dependency and constituency worlds on multiple levels, and reasons to think dependency as the more fundamental of the two ways of viewing syntax. First, the basic theories of the subclass of projective dependency grammars and phrase structure grammars both describe a set of grammars that define the class of context-free languages (Gaifman, 1965); Hays (1964) terms the theories weakly equipotent. By contrast, general, non-projective dependency graphs are strictly more expressive than constituency trees (Covington, 1990). Further, many syntactic theories combine aspects of both constituency and dependency. For example, Lexical-Functional Grammar (LFG) (Bresnan and Kaplan, 1982) and Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) augment phrase structure with grammatical functions (e.g. subject and object) and predicate-argument structures which, when viewed as connecting phrase heads, form pairwise dependencies between words. Additionally, even statistical constituency parsers make use of phrase heads internally, as heads have been found to provide a valuable addition to their probabilistic models: already ten years ago Carroll et al. (1998) wrote *"we are not aware of any contemporary parsing work which eschews the notion of head and moreover is unable to recover them."*

Finally, the immediate representation that dependency gives to grammatical functions is an obvious benefit for extracting relationships: for example, to resolve the roles of entities in an expression like e_1 *binds* e_2 it is more straightforward to follow dependencies typed *subject* and *object* than to analyse the corresponding phrase structure tree. Dependency analysis thus benefits IE regardless of the underlying theoretical framework. Ultimately, in practically oriented work it is not necessary to take a stand on whether grammatical functions are merely a derivative concept, essentially a name given to a specific configuration of the constituency tree (The Chomskyan view; see e.g. Radford, 2004, chap. 3), as their use can be motivated by the fact that they are useful. Conversely, one can make productive use of constituency-based parsers without explicitly agreeing or disagreeing with Mel'čuk (1988) that constituents only exist as a manifestation of the more fundamental dependency structure.

Dependency parsing for biomedical domain information extraction is a major topic of Papers I, II and III; the state of the art in particular is discussed in Section 4.4 of this thesis. Additionally, the specific task of parse ranking, ordering the multiple ambiguous alternative analyses returned by parsers, has been studied in our group: we introduced a parse ranking method applying the Regularized Least Squares (RLS) (see e.g. Poggio and Smale, 2003; Rifkin et al., 2003) machine learning method in (Tsivtsivadze

et al., 2005), and the topic has been considered also in a number of other studies (Pahikkala et al., 2006; Tsivtsivadze et al., 2007; Pahikkala et al., 2007; Tsivtsivadze et al., 2008). Parse ranking is discussed also in Section 3.4.

1.2.8 Relation extraction

Relation extraction[8] is the final step of the IE pipeline considered here. The task involves using the representation of the structure of the input produced in the preceding steps to identify relations of the marked entities and recognize the types of these relations. In the biomedical domain, relation extraction methods most frequently target pairwise relations, often without attempting to assign a specific relation type. The task of PPI extraction can in this case be modeled as one of deciding for each pair of proteins co-occurring in text whether or not they are stated to interact. An impressive variety of different approaches to this task have been proposed in the domain literature, but much of the variation in fundamental approach can be (somewhat informally) described by two binary choices: pattern-based vs. rule-based and hand-written vs. learned. Pattern-based approaches make use of explicit representations of expressions that state relations, while rule-based approaches only aim to decide whether or not a statement expresses a relation (cf. generative vs. discriminative models). The distinction between hand-written and learned systems is simply in whether the strategy for making a decision regarding the existence of a relation is encoded by the authors of the system or derived automatically from data.

Thus, we can differentiate between systems based on hand-written rules (e.g. Blaschke et al., 1999) and patterns (e.g. Thomas et al., 2000), systems using learned patterns (e.g. Huang et al., 2004) and, finally, those that learn decision rules (e.g. Bunescu et al., 2005). These decisions are further orthogonal to those regarding system components such as syntactic analysis: while none of the studies cited above use full parsing, it is in turn is employed, for example, by the systems proposed by Ding et al. (2003) (hand-written rule), Rinaldi et al. (2006) (hand-written patterns) Yakushiji et al. (2005) (learned patterns) and Sætre et al. (2007) (learned discriminative rule).

There are a number of reasons to favor machine learning approaches to IE. Perhaps most importantly, hand-written systems tend not to general-

[8]There is remarkable variation on the naming of this task. In a survey focusing on ACE (Doddington et al., 2004) relation extraction, Melli (2007) finds the terms *relation extraction*, *relation mention detection*, *semantic relation identification*, *semantic relation classification*, *relation detection*, *relation discovery* and *relation recognition* used by researchers studying the problem. The targets of most biomedical IE efforts can be seen as *events* (changes of entity state) in ACE terminology, but the term *event extraction* (and similar) are less frequently used in domain literature. Here the term *relation extraction* is understood broadly to include both "static" and "dynamic" extraction targets.

ize beyond the specific tasks they were designed for, and adapting a large, carefully tuned system to new tasks is a challenging and time consuming exercise requiring not only domain expertise but also detailed knowledge of the specific system (Grishman and Sundheim, 1996). Yakushiji (2006, chap. 3) argues for a learning-based approach to biomedical IE, noting as an example the *"1500 hours of highly skilled labor"* reported by Lehnert et al. (1992) as being spent on dictionary development for the MUC-3 adaptation of their CIRCUS system. While the efforts of experts are necessary also in corpus annotation for learning-based approaches, these can arguably allow a more efficient division of labor where domain experts focus only on defining the goal of extraction instead of being also intimately involved in system construction. Additionally, in a learning-based approach, IE systems in various domains can, ideally, gain the benefits of developments in machine learning through the adoption of new generic methods trained on pre-existing domain corpora.

There is a long-term trend toward learning-based approaches in IE: in the MUC conference series, which largely defined IE in the 1987–1998 period during which they were held, very few systems were learning-based (see e.g. Turmo et al., 2006), with the first fully trained system, that of Miller et al. (1998) only appearing at the last MUC event. Since then, a strong trend toward more learning-based models has emerged: Hasegawa et al. (2004) note that in the first Automatic Content Extraction (ACE) meeting (in a sense a follow-up to MUC, see Doddington et al., 2004) that introduced a relation detection task in 2002, most approaches involved learning. The same trend can be seen also in biomedical relation extraction.

While machine learning *per se* is not a major topic of Papers I–V, machine learning methods have also been studied in out group (e.g. Pahikkala et al., 2007; Pahikkala, 2008) and the design of the IE system assumes a supervised machine learning approach—learning from examples with corresponding correct outputs—to assure generalizability beyond specific domains. The viewpoint of machine learning is one of the motivations for the creation of annotated corpora and the design of the annotations as well as many other design choices in text mining systems. One of the most important issues in applied machine learning is the creation of a representation of the key features of the problem at hand for the learning machine. Such a representation should aim to explicitly include sufficient information to decide between different answers and to avoid making distinctions that make no difference to the answer (normalization). This goal motivates in part many of the design choices in a machine-learning based IE system, for example the use of dependency schemes that give immediate and systematic representation to relevant syntactic relations (see Chapter 4). Our recent efforts in constructing a kernel for learning from dependency graphs are described in (Airola et al., 2008).

The creation of general, reliable approaches to relation extraction, whether based on learning or not, is still a major challenge both outside and in the biomedical domain (Krallinger et al., 2007). The state of the art in biomedical relation extraction is discussed in detail in Chapter 6.

1.2.9 Corpora

Annotated corpora, that is, texts that have been marked up to identify, for example, named entities or syntax, are not a task in text mining in the sense discussed in the previous sections, but a central resource. Corpora are necessary for the evaluation of methods as a reference standard against which to measure their performance, as training data for machine learning methods, and for analyzing the characteristics of domain texts. Many annotated corpora have been produced in biomedical text mining studies, often created and applied in a single study and never released for wider use. However, access to the specific texts and annotations used in experiments is necessary to assure repeatability and comparability of results. Shared, publicly available corpus resources are thus a requirement for domain text mining research. Paper IV (Chapter 5) describes the BioInfer corpus and our efforts in creating this resource, and relates it to other domain resources.

1.2.10 Summary

The preceding sections have presented a brief survey of key tasks in biomedical text mining in the context of an information extraction system architecture. It was noted that for many tasks, methods with performance rivaling or matching results in general English natural language processing have been introduced. However, in the important tasks of named entity recognition, parsing and relation extraction, performance is still either unsatisfactory or (as argued further in the following chapters) to some extent unknown. The discussion of the ways in which these challenges could be addressed aimed also to motivate some of the choices guiding the research discussed in this thesis, such as the application of dependency parsing as one of the key tools and the perspective of machine learning. The above-discussed view on the state of the art in biomedical text mining and the most promising methods for addressing the main challenges has motivated the focus of work of our group, including that presented in Papers I–V. The specific aims of this research are discussed in the next section.

The above discussion of the IE tasks and the associated techniques touched on a broad range of topics, yet several related tasks, including abbreviation recognition, coreference resolution and semantic role labeling, fall out of the scope of this treatment. Additionally, this brief introduction could not cover the discussed tasks in the full detail they deserve. A more complete

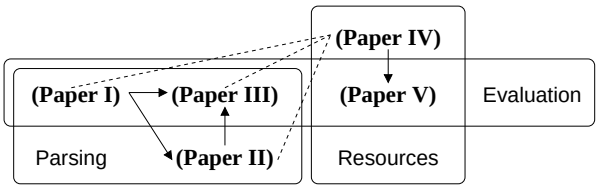

Figure 1.8: *Illustration of the relationship of the papers included in the thesis and the main topics. Arrows indicate that a study builds directly on prior work, dotted lines signal a connection through data.*

treatment can be found in the standard general NLP texts of Manning and Schütze (1999) and Jurafsky and Martin (2000), and many of the specific topics are given a chapter in the computational linguistics handbook edited by Mitkov (2003). Biomedical text mining is the topic of a recent book edited by Ananiadou and McNaught (2006) that covers the tasks discussed here and more. Surveys of different areas of the field are given by Cohen and Hersh (2005) and Spasic et al. (2005), and applications of text mining are reviewed by Ananiadou et al. (2006); references to numerous other domain surveys can be found in the recent overview of Zweigenbaum et al. (2007).

1.3 Research objectives

The overarching goals of the research of our group have largely matched those of the general biomedical text mining research program guiding numerous groups around the world: the development of techniques, methods and resources advancing the state of the art in biomedical text mining with the ultimate aims of providing useful tools for researchers working in biomedicine and developing general strategies for applying NLP methods to specific domains. Within this broad problem domain, any productive individual research must necessarily be focused on a subset of the issues. The work presented in Papers I–V and, by extension, the core of this thesis, aims to address in particular the following issues (reference to papers addressing each topic in parentheses; see also Figure 1.8):

Parsing biomedical text. One of the central aims of our study of parsers is determining the performance of available parsers on biomedical text (Papers I and III) and the identification of the challenges that general English parsers face in domain texts (Paper I). The natural follow-up to the latter question is the study of the ways in which current tools can be adapted to the domain (Paper II) as well as how their value for text mining can be otherwise improved (Paper III).

Evaluation methodology and comparability. The variety of approaches, representations, and evaluation methods used in text mining is an obstacle to determining the relative merits of different alternatives. The research presented here aimed to develop methods and techniques for meaningful evaluation (Papers I, III and V) as well as to establish the limits of comparability in the absence of standards (Paper V).

Resources for text mining. Few aspects of text mining are possible without resources such as annotated corpora. A major effort in the research presented here has been the design and annotation of such a resource for the major stages of biomedical text mining (Paper IV). The development and further refinement of resources was a key goal in our study of PPI annotations (Paper V) and a major motivating factor as well as a necessity in much of the work concerning domain parsing (Papers I, II and III).

While much of the research has been closely focused on the challenges of the biomedical domain, the approach taken has sought to avoid the trap of developing complex hand-written solutions limited to specific tasks and aimed instead for approaches that can either be applied to other tasks by training on new resources or require only little manual tuning. This approach is a recurrent minor theme of the research. Another such theme is unification, the bridging together of tools and resources that are divided on the surface but have underlying commonalities; this goal is relevant to Paper I and particularly strongly in focus in Papers III and V.

The following five chapters briefly present Papers I–V. The aim in writing these chapters has not been to summarize the entirety of the publications, but rather to introduce some of the main ideas and key results, extending the background and motivation of the research as well as relating it to relevant studies published concurrently with or after the original papers. For details, references to the publication reprints included in Part II of this thesis are included throughout.

Chapter 2

Parser evaluation

Methods for syntactic analysis, including full parsers, have been a natural part of the toolbox for biomedical text mining since early domain studies (e.g. Sekimizu et al. (1998); Craven and Kumlien (1999); Yakushiji et al. (2001)). Remarkably, while a number of domain studies had included informal or partial evaluations of the applied tools, to the best of our knowledge our evaluation (Pyysalo et al., 2004) of the Link Grammar (LG) parser was the first formal parser evaluation using a fully annotated domain corpus. Likewise, the follow-up study (Paper I) comparing the LG and Connexor Machinese Syntax (Connexor) parsers was among the first comparative domain parser evaluations, performed concurrently with, for example, the comparative study of Clegg and Shepherd (2005).

The motivation for this study is simple: parser performance is domain-dependent (Sekine, 1997), so one cannot assume that good results on general-domain text will apply to specialized domains. Without information on the performance of parsers on the domain of interest, one cannot make an informed decision on whether to apply a parser at all, nor which one to choose. A detailed evaluation can further aid in identifying specific problem areas to target in parser adaptation. We were interested in the application of dependency parsing to biomedical text mining and found a lack of information on domain parser performance, which motivated the work reported in (Pyysalo et al., 2004) and Paper I.

2.1 Parsers

In Paper I, we evaluated and compared the LG parser of Sleator and Temperley (1991) and the commercial Connexor parser based on the Constraint Grammar framework (Karlsson, 1990). LG and Connexor represent two fundamentally different approaches to parsing (Paper I, Section 2), but share the important property of producing dependency-type parses.

The specific choice of LG and Connexor as the targets of the evaluation reflects both the desired practical advantages of dependency representations for text mining as well as the relative popularity of these methods in biomedical domain studies. LG had received considerable attention in the biomedical text mining literature: among the earlier studies, Ding et al. (2003) used the parser for PPI extraction with a simple decision rule based on parse connectedness, Phuong et al. (2003) introduced a method for learning patterns expressing interactions using LG parses, Aubin (2003) presented a parser comparison including LG, Szolovits (2003) introduced an automated extension of the parser lexicon with terms from the UMLS Specialist lexicon (see Chapter 3 of this thesis), and the Caderige project (Alphonse et al., 2004) applied LG for domain ontology population as well as PPI extraction.

Likewise, methods based on the Constraint Grammar framework have been frequently applied to biomedical domain tasks. EngCG (Voutilainen, 1997) was applied by Sekimizu et al. (1998) in a rule-based PPI extraction system and later by the same group for reducing lexical ambiguity prior to full parsing (Yakushiji et al., 2001). Connexor was similarly applied in the Yapex named entity recognition system of Franzén et al. (2002), for initial predicate-argument structure annotation by Wattarujeekrit et al. (2004), and in the PPI extraction systems of Koike et al. (2003, 2005).

2.2 Evaluation methodology

A large number of methods and approaches to parser evaluation have been proposed (see e.g. Carroll et al., 1998; Kakkonen, 2007, chap. 8), of which methods employing annotated corpora are a widely used and reliable way for establishing performance. Within the dominant constituency framework, standard measures proposed by Black et al. (1991), known as the PARSEVAL (or GEIG) measures, compare parse trees in terms of representations where phrase boundaries are marked by brackets. The quality of a parse with respect to the correct parse is evaluated in terms of the number of crossing brackets and precision and recall of the correct phrases. While the PARSEVAL measures are standard and still widely used, they have faced much criticism: one simple, specific objection is that a perfect score can be achieved on the crossing-brackets measure by assigning the sentence no structure at all. Magerman judged the crossing-brackets measure *"misguided"* and the measures in general *"crude"*, arguing instead for exact matching of all properties of the entire tree (Magerman, 1994, chap. 7). Lin (1995) argued similarly that the measures count some errors multiple times and do not accurately reflect how much a parse deviates from being correct, suggesting instead dependency-based evaluation using an error-count.

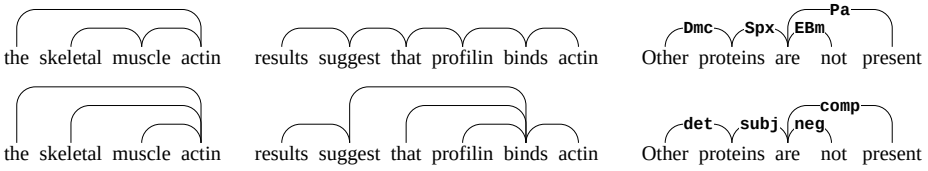

Figure 2.1: *Examples of differences between the LG (above) and Connexor (below) dependency schemes. [Figures from Paper I]*

Dependency-based parser evaluation has recently found general support (e.g. Carroll et al., 1998; Preiss, 2003; Kaplan et al., 2004; Clark and Curran, 2007; Nivre et al., 2007), as discussed in more detail in Chapter 4 of this thesis. Clegg (2008) (Chapter 2) considers a number of constituency and dependency-based evaluation methods and presents a recent, detailed argument in favor of dependency for parser evaluation. Clegg's endorsement of dependency-based evaluation is particularly relevant here as it is informed by his two recent biomedical domain studies on parser evaluation, the first based on constituency and the latter on dependency (Clegg and Shepherd, 2005, 2007). The dependency-based evaluation in Paper I is thus not only a natural choice for dependency parsers, but also in line with current practice in parser evaluation.

Evaluation of parser performance in terms of dependency relations is not sufficient alone to guarantee meaningful comparability of results for different parsers. The dependency schemes applied by parsers differ with respect to the recognition of heads and dependents, the types assigned to different dependencies, and the number of dependencies and dependency types employed: Figure 2.1 illustrates some differences between the LG and Connexor schemes. Because of these differences, it is not possible to evaluate the parsers on a single set of annotations. One alternative would be to create two separate dependency annotations, each following in detail the scheme of one of the parsers. However, there is no guarantee that the results of such separate formalism-specific evaluations can be meaningfully compared (see e.g. Miyao et al., 2007). One way to ensure comparability is to create conversions from parser-specific representations into a common scheme. As discussed in detail in Paper III (Chapter 4 of this thesis), there are several challenges in making this approach work, relating in particular to large margins of error from the conversion.

In Paper I, to create a meaningful comparative evaluation we evaluated each parser on an annotated corpus (the *gold standard*) that is essentially specific to its dependency scheme, yet abstracted from the largest differentiating factors between the two schemes. We further aimed to guarantee that the number of constraints on their output that the two parsers must meet

are as close to equal as possible. Specifically, to ensure comparability, we evaluated each parser against a separately annotated gold standard that follows its dependency scheme structurally, but we did not require dependency types to match: LG uses approximately 10 times as many specific dependency types as Connexor, so requiring types to match would have placed unequal demands on the parsers. Additionally, as LG dependencies are not explicitly directed, we only required that the correct words be connected. Finally, we modified the LG gold standard scheme to exclude a number of dependencies that form cycles in the original LG scheme, so that the dependency graphs for each parser would form trees. The resulting gold standards contain a total of 7541 dependencies in the modified LG scheme and 7540 in the Connexor scheme, thus meeting the aim of having close to the same number of constraints for each. The selection of the gold standard sentences and their annotation is discussed in Section 4 of Paper I.

2.3 Evaluation criteria

Dependency evaluation is typically performed in terms of precision and recall of (specific types of) dependencies, summarized using the F-measure. Due to the exclusion of some of the dependencies in the gold standard annotation, we could not measure precision: there was no way to automatically determine whether an additional dependency is correct. Generally, measurement of recall without precision is close to meaningless: a fully connected graph would trivially maximize the metric. However, in comparisons of connected, acyclic dependency trees, precision equals recall (and hence F-measure): there will be exactly one extra dependency for each missing dependency. While neither the gold standard nor the parser outputs are always acyclic and connected, exceptions are rare. The assumption that recall can in this case be used as a meaningful proxy for F-measure performance was partly empirically validated by the LG-specific evaluation in Paper III (Table 2, LG scheme), where both precision and recall were measured and differed only by one percentage unit.

The fraction of gold standard dependencies that each parser recovers provides an intrinsic view of parser performance. To gain insight into the applicability of the parsers specifically to PPI extraction, we introduced the *interaction subgraph* concept for parser evaluation, measuring the recovery of complete subgraphs connecting two entities in a way that states an interaction. The basic idea of giving preferential status to the minimal substructure of a syntactic representation that contains two entities whose possible relation is being studied is a natural one, and has been frequently applied, reinvented and explored in IE studies (e.g. Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Zhang et al., 2006;

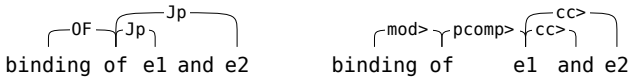

Figure 2.2: *Example where the shortest path between e_1 and e_2 excludes the word stating the interaction. LG representation on the left, Connexor on the right.*

Fundel et al., 2007; Erkan et al., 2007; Kim et al., 2008c). The core intuition is formalized by Bunescu and Mooney (2005) as the *Shortest Path Hypothesis*:

> *"[for two entities e_1 and e_2] the contribution of the sentence dependency graph to establishing the relationship $R(e_1, e_2)$ is almost exclusively concentrated in the shortest path between e_1 and e_2 in the undirected version of the dependency graph."*

(see also Bunescu, 2007, sec. 4.4.1). This hypothesis holds to varying degrees for different dependency schemes. Counterexamples are easy to find for LG and Connexor, among others (Figure 2.2). To capture the relevant information in such and in more complex cases, we defined the interaction subgraph as follows:

> *The interaction subgraph for an interaction between two proteins A and B in a dependency parse $\mathcal{P}$ is the minimal connected subgraph of $\mathcal{P}$ that contains A, B, and the word or phrase that states their interaction.* [Paper I]

Interaction subgraphs capture the relevant information regarding PPIs more accurately than shortest paths in a number of cases. Nevertheless, as correctly pointed out by Clegg and Shepherd (2007), the interaction subgraph may exclude words such as negations and is thus also only an approximation of the information that needs to be considered to extract PPIs[1].

Finally, in addition to measuring the recovered dependencies and interaction subgraphs, we determined the fraction of fully correct parses, that is, parses for which all dependencies were correctly recovered. This measure is analogous to the exact match measure of (Magerman, 1994, chap. 7) and allows the results to be related to evaluations where the fraction of fully correctly parsed sentences is employed as a metric.

Criterion	LG (first)	LG (best)	Connexor
Dependency	72.9%	81.3%	80.0%
Fully correct	7.0%	27.0%	7.0%
Interaction	26.9%	58.1%	36.2%

Table 2.1: Parser performance. [From Paper I, Tables 1 and 2]

2.4 Results

The main results of the evaluation are summarized in Table 2.1. While the Connexor parser only returns one analysis for each sentence, the LG parser returns all ambiguous alternative analyses allowed by the grammar, and performance for LG is given separately for the first parse, that is, the parse ranked highest by the parser heuristics, and the best parse out of the given alternatives ("oracle" ranking). For the most directly comparable result, the dependency-level performance for the first parse of LG against Connexor, the latter performs statistically significantly better (see Paper I for details), with an error rate 30% lower than that of LG. To relate these numbers to performance on general English, we noted that an evaluation of the Functional Dependency Grammar parser on which Connexor builds has been reported to achieve over 88% performance on a variety of genres (Tapanainen and Järvinen, 1997). While several caveats regarding comparability apply, this suggests that the error rate of the parser may increase by over 60% when moving to biomedical text. The results for fully correct parses and to a lesser extent interaction subgraphs are very low, suggesting that both parsers face serious problems on biomedical domain text and indicating a need for domain adaptation. Further, the considerably better result for the best parses of LG compared to those ordered first by parser heuristics indicates an opportunity to improve performance by reranking the parses.

The fully correct, best parse result for LG (27%) is comparable to the best estimated result (31%) achieved for a similar metric in the adaptive evaluation of Grover and Lascarides (2001) of the Alvey Natural Language Tools (ANLT) parser (Grover et al., 1993) on the unannotated medical-domain OHSUMED corpus (Hersh et al., 1994). While the dramatically poorer initial result for ANLT (2%) suggests that the robust parsing algorithm of LG (Grinberg et al., 1995) allows it to overcome some of the issues causing ANLT to fail, both results call into question the applicability of parsers employing full hand-crafted general English grammars to the

[1]While simple negations as in *A does not bind B* can be resolved with special-case rules such as those applied by Bunescu and Mooney (2005), the general case is harder; consider, for example, *Our experiments failed to find any support for the hypothesis that A binds B.* (see e.g. Sanchez-Graillet and Poesio, 2007)

biomedical domain. Rebholz-Schuhmann et al. (2005) ask if text mining is ready to deliver and, referring to these results, conclude that *"computational linguists have not yet developed tools that can analyse more than 30% of English sentences correctly."* We would not generalize these results quite as broadly, as they do not represent, for example, the performance of statistical parsers (see Chapter 4 in this thesis). Further, as Rebholz-Schuhmann et al. accurately note, for the extraction of a particular fact one need only correctly analyse part of the sentence, specifically that stating the fact. Of the results in Table 2.1, we would thus emphasize the somewhat more promising numbers for interaction subgraphs.

As part of this study, we further performed a detailed failure analysis of LG, with suggestions on how the failures could be addressed (Paper I, Section 7) and an evaluation of the effect of two approaches to adaptation (Paper I, Section 8). As these topics relate closely to domain adaptation, discussion will be deferred to Chapter 3 of this thesis, which is focused on the task. However, a criticism of one aspect of the work in Paper I (as originally presented by Pyysalo et al., 2004), should be addressed here. McNaught and Black (2006) (pages 154–156) provide a detailed and largely very positive discussion of the study, but criticize our estimate that 8% of failures are due to ungrammatical sentences and suggest that we may be observing natural sublanguage behavior that only appears ungrammatical. First, there are cases in the corpus that are genuinely ungrammatical, such as agreement errors

- *Reduced expressions of cell adhesion molecules (E-cadherin, alpha-catenin, and beta-catenin) has been reported [...]*

- *The high recombination levels seen in rad5 and rad18 mutants is dependent on [...]*

as well as some borderline cases such as missing coordinating conjunctions

- *The contents of myosin heavy chain, myosin light chain 2, actin, troponin-I in 125-week-old rats decreased [...].*

However, the core of their observation is accurate, as the most frequent cause of failure categorized as ungrammatical, dropping otherwise mandatory determiners (Paper I, Section 7.1), can be seen as accepted sublanguage usage, in particular in publication titles. Thus, while these sentences are ungrammatical from the point of view of the LG general English grammar—which is why the term "ungrammatical" was selected—this term was perhaps incautiously generally applied to describe the whole of this category of reasons for parser failure.

2.5 Conclusions

The work presented in Paper I was an early effort toward answering currently more thoroughly studied questions (see Chapter 4 of this thesis) regarding the performance of full parsers on biomedical text with reference to fully annotated domain corpora. The evaluation indicated that both the LG and Connexor parsers have problems on domain text, further characterizing these problems in detail for LG. The following chapter extends on these topics, focusing on the important related issue of domain adaptation.

Chapter 3

Domain adaptation

It has long been recognized that language use varies substantially by domain, with sublanguages such as that of biomedical research being characterized by specialized vocabulary, syntax, and semantics (Harris, 1968; Grishman, 2001; Friedman et al., 2002). Differences in lexicon and syntax, in particular, contribute to the domain dependence of parsing (Sekine, 1997). The relatively low performance of parsers in the domain (Paper I) serves as a motivation for specific efforts to adapt parsers to biomedical texts. In Paper II, in collaboration with Sophie Aubin and Adeline Nazarenko of Université Paris-Nord, we applied the evaluation methodology and an extended version of the corpus of Paper I to study domain adaptation methods. The interest in the Link Grammar (LG) parser in domain studies and the fact that the parser is an open system that can be freely modified according to need served to motivate the choice of LG for studying adaptation methods.

The findings of Paper I as well as those reported by Aubin et al. (2005) indicated that issues relating to ambiguity and the lexicon are prominent among the problems that LG faces in parsing biomedical English. In the work described in Paper II, we chose to concentrate in particular on lexical adaptation, that is, adaptation addressing domain vocabulary, including unknown words, a major source of ambiguity. In addition to the indication that lexical issues are a particular problem area for LG, our choice of focus addresses also the general aim that adaptation methods should not require intensive manual efforts. While changes to the extensive hand-written grammar of the LG parser were also known to be necessary to improve performance, such changes would, for the most part, only benefit the specific subdomain that the adaptation targets. By contrast, by studying lexical adaptation approaches that require little or no manual work, we aimed not only to extend the applicability of the parser but also to establish the benefits and limits of such adaptation in a way that would generalize also to other domains.

3.1 Adaptation methods

We chose to consider three lexical adaptation methods, two of which were proposed in the recent literature on domain parsing: the automatic mapping of terminology between lexicons proposed by Szolovits (2003) and the surface feature-based lexical disambiguation approach proposed by Aubin et al. (2005). We additionally chose to evaluate the effect of using part-of-speech taggers to disambiguate unknown words prior to parsing. These methods are briefly presented next; more details are included in Paper II.

The method of Szolovits (2003) maps information between lexicons as follows: to determine the lexical description that should be given to a word w that is found in the source lexicon S but not in the target lexicon T, the method determines first the set of words $S_w \subseteq S$ that share the same lexical description as w and then the subset of those words that also occur in T, $S_w \cap T$. One of the descriptions of the words in $S_w \cap T$ is then assigned to w based on the overlap between these sets. Szolovits applied this method to augment the LG lexicon with words from the Unified Medical Language System (UMLS) Specialist lexicon (McCray et al., 1993; Bodenreider, 2004), adding over 100,000 words. This addition more than tripled the number of words defined for LG and was shown to considerably extend coverage of words in a clinical domain corpus. This large extension based on a major lexicon covering a closely related domain presented a good candidate for lexical adaptation. However, the focus of UMLS Specialist is on medical, not biological terminology, suggesting that the increase in coverage might not extend to text discussing, for example, protein-protein interactions. Further, as no study on the effect on parse quality was presented by Szolovits, it was necessary to evaluate the extension prior to adopting it.

Aubin et al. (2005) studied the requirements of adapting the LG parser to the biomedical sublanguage and suggested a number of modifications, including normalizing preprocessing steps, lexicon extension, incorporation of term recognition, and the modification of grammar rules. They additionally identified a number of largely domain-specific suffix morphological rules (see e.g. Mikheev, 1997) that could be used to disambiguate unknown words by extending the "morpho-guessing" system used in standard LG lexical processing. The identification of domain-specific surface features that imply specific lexical categories can be performed by analyzing unannotated texts, and does not presuppose the existence of large-scale lexical resources such as UMLS or the availability of POS-annotated corpus resources. This rule-based disambiguation method thus represents a relatively lightweight, general approach to domain adaptation.

In the study originally proposing the use of morphological features for biomedical domain adaptation Aubin et al. (2005) proposed 19 disambiguation rules; this number was extended to 23 based on further analysis for the

Suffix	Part of speech	examples
-ase	noun	synthetase, kinase
-ity	noun	chronicity, hypochromicity
-on	noun	replicon, intron
-ose	noun	isomaltotetraose, isomaltotriose
-yl	noun	hydroxyethyl, hydroxymethyl
-ide	noun	iodide, oligodeoxynucleotide
-ic	adjective	glycolytic, ribonucleic, uronic
-ive	adjective	nonpermissive, thermosensitive
-ble	adjective	inducible, metastable
-ae	adjective (latin)	influenzae, tarentolae
-um	adjective (latin)	japonicum, tabacum, xylinum
-fold	adjective/adverb	10-fold, 4.5-fold, five-fold

Table 3.1: Examples of morphological features used for disambiguation. [From Paper II, Table 1]

evaluation presented in Paper II. A number of these rules, the general classes of words the features imply, and some examples to which the rules apply are shown in Table 3.1. It should be noted that the rules are intended to apply to unknown words and rely on the existence of a broad-coverage general language lexicon: for example, the word *increase* would only be recognized as a noun by these rules if it were not found in the lexicon.

High-reliability biomedical domain part-of-speech taggers (e.g. Smith et al., 2004; Tsuruoka et al., 2005) had recently become available at the time of the study. The training of accurate machine learning-based POS taggers requires a considerable amount of training data: the version of the GENIA corpus used by Tsuruoka et al. (2005) contains almost 500,000 manually tagged words. This approach is thus not as readily applicable to lexical adaptation to new domains where resources are scarce as the other two adaptation methods. Nevertheless, POS annotation is less demanding to create than the full syntactic annotation required for the retraining of statistical parsers. Additionally, domain POS taggers are important tools and their effect on parsing performance is both interesting as a separate piece of information and a relevant point of comparison.

The incorporation of POS information required two modifications of the parser: a way to pass POS-tagged words as input and a way to map these tags into the lexical descriptions used by the parser. These changes were implemented for the study; examples of the introduced mappings from POS tags to LG lexicon entries are shown in Table 3.2.

Tag	Description of tag	LG rule	Description of LG rule class
NN	common noun, sing.	words.n.4	sing. nouns, mass or count
NNS	common noun, pl.	words.n.2.s	plural nouns
JJR	adjective, comparat.	words.adj.2	comparative adjectives
JJS	adjective, superlat.	words.adj.3	superlative adjectives
VB	verb, base	words.v.6.1	optionally transitive verbs
CD	number	NUMBERS	general rule for numbers
RB	adverb, base	words.adv.1	ordinary manner adverbs

Table 3.2: Examples of POS tag mapping to LG rules. [From Paper II, Table 2]

3.2 Evaluation methodology

Evaluation focused on three aspects of the effect of the adaptation methods: vocabulary coverage, ambiguity, and parsing performance. Vocabulary coverage was evaluated on two corpora, in Paper II called *interaction* and *transcript*. The first is an extension of the annotated corpus used in Paper I and represents a subset of the syntactic annotation contained in the BioInfer corpus (Paper IV), and the latter is a 17,000 sentence unannotated corpus on the *Bacillus subtitlis transcription* subdomain. Ambiguity and parsing performance were evaluated on the annotated corpus.

Vocabulary coverage was measured by the fraction of corpus words that are found in the lexicon, match a surface feature disambiguation rule, are disambiguated based on POS, or are unknown. The modified LG applies these alternatives in this order as a cascade—the first that matches determines how a word is processed. Ambiguity was measured by average sentence parsing time and the average number of analyses returned per sentence. Finally, the evaluation of parsing performance followed the approach described in Paper I (Section 2.2 of this thesis).

3.3 Results

Figure 3.1 illustrates the frequency with which different methods apply to assign lexical descriptions to tokens in the unmodified parser (Orig) and the adaptations with the UMLS Specialist lexicon extension, the morphological feature extension (xMG) and the POS extension. One interesting observation is that the fraction of remaining unknown words is not considerably larger for the xMG extension that adds just 23 simple rules than for the 100,000 word lexicon extension. The generality of these rules is supported by the fact that they match equally on the interaction corpus, which represents a subdomain that was not considered in rule development.

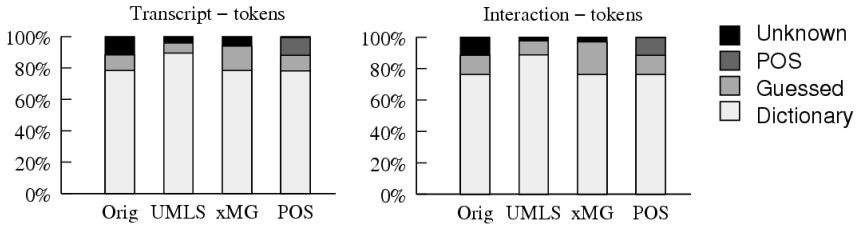

Figure 3.1: *Fraction of running tokens covered by each of the four methods on the two corpora. [From Paper II, Figure 2]*

The effect on ambiguity is shown in Table 3.3. While the number of different analyses is decreased by about a third for all adaptations, the effect on parsing time diverges somewhat, possibly reflecting in part the specificity of the assigned lexical descriptions.

Metric	Orig	UMLS	xMG	POS
Time	15.4s	9.9s	10.8s	8.6s
Lkg. ratio	1	0.67	0.68	0.66

Table 3.3: Ambiguity with different adaptations. Time is average parsing time per sentence, linkage ratio is average of per-sentence linkage number ratios. POS results using GENIA tagger. [From Paper II, Table 3]

Finally, the effect on performance is shown in Table 3.4. The reduction in error rate for the xMG and POS approaches is encouraging, but does not bring the results to the level expected of accurate parsers on general English. The best observed relative decrease in error of 10% provides an indication of the improvement that can be achieved through this type of lexical adaptation.

	Orig	UMLS	Δ	xMG	Δ	POS	Δ
recall	74.2	75.4	4.7	76.0	7.0	76.8	10.1
p	N/A	$p \approx 0.06$		$p < 0.01$		$p < 0.01$	

Table 3.4: Parser performance with different adaptations. Recall is given for the parse ordered first by the parser heuristics, Δ columns give relative decrease in error with respect to the original LG, and p values are estimates of the significance of this difference (signed-ranks test; Wilcoxon, 1945). [From Paper II, Table 4]

These results agree with the findings of the failure analysis performed in Paper I and confirm that modification of the hand-written grammar of the parser is also necessary in domain adaptation.

Regarding the implications of these results on the relative merits of the adaptation methods, the result that performance is best with an accurate domain tagger was expected, but the finding that the simple rule-based model gives better performance than the large lexicon extension is somewhat surprising. In Paper II we analyse this in detail, showing that the lexicon extension can in some cases lead to reduced parse quality due to mapping errors. The competitive performance of the rule-based method encouraged us to apply a similar method to unknown word resolution in our recent work on parsing Finnish (Laippala et al., 2008).

As part of the study we performed also experiments with the popular Brill POS tagger (Brill, 1992) trained on general English texts (used in biomedical text mining at least by Ono et al., 2001; Huang et al., 2004; Hao et al., 2005), finding a high error rate and a mixed effect on performance. We further evaluated a number of additional aspects of the effect of the adaptations on performance and tested all possible combinations of the adaptations, but found no notable further benefit. These experiments are discussed in detail in Paper II.

3.4 Discussion and Conclusions

The experiments in Paper II supported the value of an accurate domain POS tagger for the lexical domain adaptation of a general parser, finding among other benefits a 10.1% reduction in error. As there is no theoretical reason to expect this result to generalize to parsers based on fundamentally different approaches from that of LG, it is interesting to note the similarity to results reported by Lease and Charniak (2005) for the domain adaptation of a statistical treebank parser (Charniak, 2000): Lease and Charniak found an 11.5% reduction in error from POS-based adaptation, remarkably close to that observed here. In addition to the three adaptation approaches considered in Paper II, as part of the study reported in Paper I we considered also the effect of employing "oracle" knowledge of named entities in support of parsing, which can be seen as a form of lexical adaptation. In this experiment we found a 16.8% relative decrease in error; Lease and Charniak considered the same strategy, reporting a comparable 12.0% decrease. The similarity of these results across parser formalism and evaluation strategy suggests that they may be tentatively considered as indicative of the general effect of such adaptations.

In Paper I we observed that there is a considerable difference in quality between the parse ranked first by the LG heuristics and the best parse among the ambiguous alternatives allowed by the grammar, suggesting that reranking the parses could improve performance. We investigated this possibility in a separate study (Tsivtsivadze et al., 2005), finding that training on

as few as 500 annotated in-domain sentences could provide a 9.5% relative reduction in error. While integrating such a reranking method would thus further improve parser performance, this is not solely an effect of domain adaptation as the original LG parser contains no statistical ranking component. Related results from training on fully annotated domain corpora are considered further in the discussion of the current state of the art in domain parsing in Section 4.4 of this thesis; domain adaptation has also been recently studied from a different perspective in a shared task in the Computational Natural Language Learning (CoNLL'07) conference (Nivre et al., 2007).

As part of the work described in Paper II we created an adapted version of the LG parser that offers a 45% decrease in parsing time and a 10% relative decrease in error over the unmodified LG when used together with a domain tagger. This parser was made freely available as the first release of BioLG. It has since been further developed to include a number of modifications to its grammar, as proposed in Paper I and (Aubin et al., 2005). The parser has also been incorporated into the Alvis NLP platform (Deriviere et al., 2006) and applied in a number of studies, including that described in Paper III.

Chapter 4

Unifying syntactic representations

Parsing technologies have long been fragmented by the different formalisms employed. The largest divide is between dependency and constituency, but even parsers based on more closely related formalisms employ substantially different syntactic representations. This has a number of unfortunate consequences: corpora tend to be formalism-specific, reducing the amount of data practically available, parser evaluations yield results that cannot be directly compared, and methods that apply parsers tend to become bound to the formalism, making it difficult to select between different tools.

There has recently been considerable interest in unification using standard syntactic representations. Three prominent proposals are dependency-based: the Grammatical Relations (GR) scheme proposed by Carroll et al. (1998), the scheme used in the 2006–2007 CoNLL dependency parsing tasks (Nivre et al., 2007) and the Stanford dependency scheme (SD) of de Marneffe et al. (2006). The GR and CoNLL schemes are intended for parser evaluation[1], while the SD scheme, although originally inspired by GR, aims in particular to be useful for further applications. The Stanford parser package also includes a conversion from the standard Penn Treebank (PTB) constituency representation to the SD scheme. The trend toward unification and the advantages of shared, application-oriented syntactic representations provided the motivation for work described in Paper III.

[1]Carroll et al. (1998) explicitly wrote that they are not advocating GR for use in application tasks, although more recent work appears to partially reverse this stance (Briscoe and Carroll, 2002).

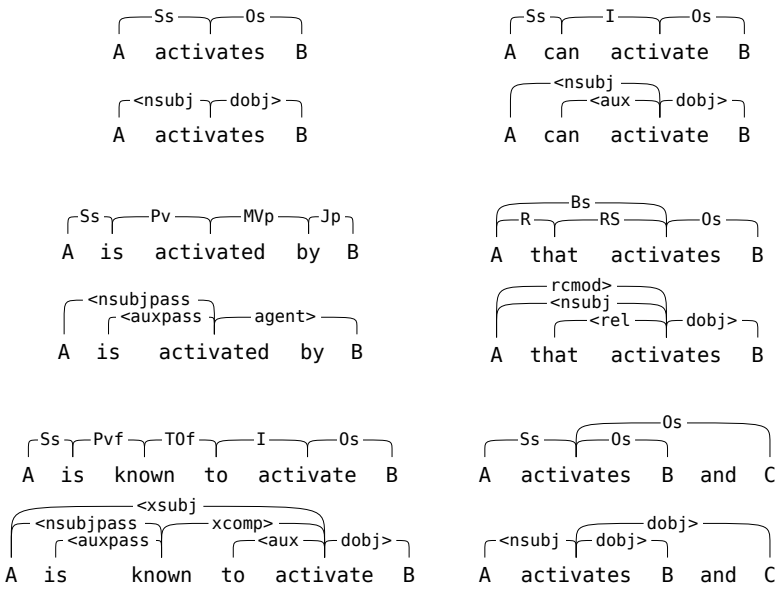

Figure 4.1: *Illustration of LG and SD representations. The syntactic structures assigned in the LG (above) and SD (below) representations to cases exhibiting selected grammatical phenomena.*

4.1 Dependency representations

One of the key benefits of performing syntactic analysis as part of an information extraction pipeline is normalization, abstracting away unnecessary or coincidental differences between different forms of expression carrying the same (or similar) meaning. From the perspective of normalization, the output representations that parsers use can make a great difference to their value for further processing. Unlike the SD scheme, the dependency scheme employed by the LG parser was not designed with IE applications in mind, and for this purpose its dependency types in part make unnecessary distinctions and in part lack relevant distinctions. Moreover, LG dependency structures are oriented toward surface syntax rather than semantics. Figure 4.1 shows the syntactic structures for a number of expressions in the LG and SD schemes, illustrating that the SD representation more directly represents the semantically relevant connections and offers a more consistent representation of the underlying structure in the face of surface variation (see also Paper III, Section 3).

Our recent study comparing different dependency representations against a representation of biomedical relations as semantic dependencies indicates that the "collapsed" SD representation in particular closely captures the relevant semantics (Björne et al., 2008). Related issues are also discussed

by de Marneffe et al. (2006) and by Yakushiji (2006) (pages 7–10) in the context of the deep parser Enju (Miyao and Tsujii, 2005). Miyao et al. (2008) recently performed a PPI task-oriented parser evaluation that lends further empirical support to the value of dependency-based representations in general and SD in particular: excepting Enju, the evaluated parsers consistently performed best when their output was converted to SD or the CoNLL (Nivre et al., 2007) dependency representation, even though their native PTB output is constituency-based.

The introduction of the SD scheme and the accompanying conversion from the PTB scheme made many constituency parsers and corpora easily applicable in application-oriented systems that use dependency representations, combining the benefits of a large treebank and a semantically oriented output representation. While this conversion is far from the first to identify head-dependent relations in constituency trees (see e.g. Magerman, 1994, Section 5.2.2. and Appendix C) nor the first to specifically extract typed semantically-oriented dependencies from PTB trees (Levy and Manning, 2004), it has, in addition to these properties, the benefits of being available in a standard software package and easily applied. A further point in its favor here is its recent use in the detailed evaluation of several statistical parsers on the biomedical GENIA treebank corpus (GTB) by Clegg and Shepherd (2007).

Recognizing both the value of unification under shared representations and the potential benefits of the SD representation for further applications, we created a conversion from the LG to the SD scheme as part of the work presented in Paper III. A reliable conversion provides a number of benefits, increasing the value for IE applications of both the LG and BioLG parsers as well as that of the syntactic annotation of the BioInfer corpus, originally created using the LG scheme (Paper IV, Chapter 5 of this thesis). Together with the PTB→SD conversion, a LG→SD conversion also makes it possible to use the syntactic annotations of the PTB-annotated GTB and BioInfer corpora together, and further allows the performance of statistical treebank parsers, LG and BioLG to be evaluated on a shared scheme.

4.2 Conversion methodology

To create the LG→SD conversion, we used a combination of preprocessing and hand-written rules. First, preprocessing was applied to resolve coordination, which is analysed in fundamentally different ways in the LG and SD schemes. Conversion rules were written using the lp2lp formalism developed by Alphonse et al. (2004)—lp2lp had previously been applied by Aubin (2005) to convert parses initially generated by LG into a more application-oriented scheme for the LLL challenge (Nédellec, 2005).

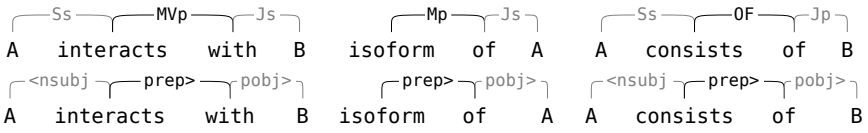

Figure 4.2: *Illustration of conversion rule. LG structures above, SD below. LG dependencies matching the regular expression* Mp|MVp|OF *and their corresponding SD* prep *dependencies highlighted.*

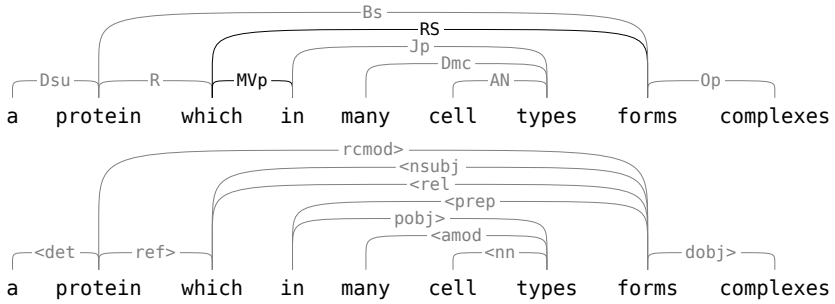

Figure 4.3: *Illustration of conversion rule. LG structure above, SD below. LG* MVp *and* RS *dependencies highlighted.*

The conversion rules consist of constraints on the source LG parse, and most typically specify that a dependency must be found for the rule to apply. In addition to the dependencies, constraints can refer to the lexical level, requiring certain words to be found in the sentence. When the constraints of a rule are satisfied, one dependency is generated to the target parse, with the head, dependent and dependency type specified by the rule. As an example of the simplest possible form or rule specifying only one dependency type constraint, $[\mathrm{X} \overset{\mathrm{Pv}}{\text{—}} \mathrm{Y}] \Rightarrow \mathrm{Y} \overset{\mathrm{auxpass}}{\rightarrow} \mathrm{X}$ specifies that whenever a dependency of type *Pv* is found in the source connecting any two words X and Y, an *auxpass* dependency where Y is the head and X the dependent is generated. Further, constraints can be negated, requiring that a dependency is not present, and dependency types can be specified with regular expressions, as in the following example (simplified from Paper III):

$$[\mathrm{X} \overset{\mathrm{Mp|MVp|OF}}{\text{—}} \mathrm{Y}]![\mathrm{X} \overset{\mathrm{RS}}{\text{—}} \mathrm{Z}] \Rightarrow \mathrm{X} \overset{\mathrm{prep}}{\rightarrow} \mathrm{Y}$$

Figure 4.2 illustrates three cases where the constraints of this rule are satisfied and a *prep* dependency generated; Figure 4.3 illustrates a case where the negated constraint matches and no dependency is generated by this rule. Discussion of the preprocessing as well as a more detailed description of the rules is found in Paper III, Section 4.1.

It should be noted that the decision to create the conversion rules by hand goes somewhat against the general aim of avoiding manual work whose results do not generalize. However, to reach the general goal of unifying syntactic representations under a single shared representation, it would suffice to generate as many conversions as there are different syntactic schemes. Compared to the development of the parsers for the various schemes, the development of conversion rules—in this study estimated as taking approximately 100 hours—is far from an excessive effort[2]. Further, the conversion rules do not contain any domain-specific aspects and thus require no adaptation when applying parsers to novel tasks. As there are no established methods for creating high-quality conversions between dependency schemes, we opted to create the rules by hand as the most reliable alternative. If large numbers of conversions need to be created, the development of learning methods for automating the process would become a valuable alternative approach.

4.3 Results

Rule development was performed with repeated testing against a partial reference standard until a point of clearly diminished returns for effort was reached. The created ruleset consisted of 114 rules, each rule specifying on average 4.4 constraints (for more detailed statistics, see Paper III, Section 5). The rules were then applied to tentatively convert the BioInfer corpus gold standard LG annotation into the SD scheme, and the resulting annotation was then manually corrected so that each sentence was separately corrected by two independent annotators. Inter-annotator agreement was measured as 97.4% F-measure, a high result supporting the stability of the SD scheme; all remaining disagreements were resolved jointly by the annotators.

To evaluate the rules, they were used to convert the LG scheme gold standard annotation of the corpus and the result compared against the created gold standard SD scheme annotation. Figure 4.4 illustrates the cumulative precision and recall of the conversion when rules are added in highest-recall-first order. The full ruleset reaches a very high precision of 98.0% and recall of 96.2%, indicating that the creation of a high-reliability conversion is possible. The quality of the conversion was further ascertained by evaluating the BioLG parser on the LG and SD versions of the BioInfer corpus syntactic annotation, finding no loss in performance from the conversion. This result is discussed further in Paper III, Section 5.3.

Finally, we applied the newly created SD BioInfer and an SD-converted version of the GTB corpus to evaluate the BioLG parser and the version

[2] A considerably larger effort was expended to create a system supporting rule writing and quality measurement, but this exercise need not be repeated for other rulesets.

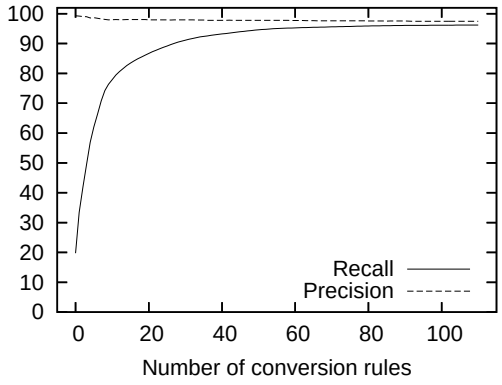

Figure 4.4: *Cumulative precision and recall of the LG→SD conversion rules. [Figure from Paper III]*

of the Charniak statistical treebank parser (Charniak, 2000) adapted to the biomedical domain by Lease and Charniak (2005) (below Charniak-Lease, or C-L), which performed best overall out of the nine variants of five parsers evaluated by Clegg and Shepherd (2007). The results of this evaluation are shown in Table 4.1[3]

	Charniak-Lease			BioLG		
corpus	Prec.	Rec.	F	Prec.	Rec.	F
GENIA	81.2	81.3	81.3	76.9	72.4	74.6
BioInfer	78.4	79.9	79.1	79.6	76.1	77.8

Table 4.1: Parser performance. Precision, recall and F-measure for the two parsers on the two corpora. [From Paper III]

Despite an approximately 5% divergence on the measured F-measure performance difference between the parsers on the two corpora due to conversion biases (discussed in Paper III, Sections 5), the C-L parser achieves statistically significantly better performance on both corpora, demonstrating the feasibility of using the SD scheme as a unifying representation for parser comparison.

4.4 Discussion and conclusions

The study presented in Paper III is the most recent of our parsing-related work included in this thesis, and this chapter thus the last dedicated to

[3]The table presented here corrects a slip of the finger in data entry for the table in Paper III, where the figure 79.4 appears as the F-measure of the C-L parser on BioInfer instead of the correct 79.1. This difference does not alter the conclusions of the study.

related topics here. The following brief discussion covers related work published concurrently with or after Papers I to III to further set these studies in their broader context.

study	source	target	precision	recall	F-measure
Clark and Curran (2007)	CCG	GR	86.9%	82.8%	84.8%
Paper III	LG	SD	98.0%	96.2%	97.1%
Sagae et al. (2008a)	HPSG	GR	87.5%	86.8%	87.1%
Sagae et al. (2008a)	SD	GR	80.8%	69.2%	74.5%
Sagae et al. (2008a)	HPSG	PTB	98.1%	98.1%	98.1%
Haverinen et al. (2008)	Pro3Gres	SD	96.9%	95.4%	96.1%

Table 4.2: Reported conversions with conversion quality estimates.

First, there is currently considerable interest in formalism-independent parser evaluation both in general and biomedical domain NLP, and the related issue of conversion between syntactic representations has lately been the focus of a number of studies. The GR representation of Carroll et al. (1998) can be seen as emerging as a standard for formalism-independent evaluation of parser performance, while SD (de Marneffe et al., 2006) is gaining popularity in IE applications and used in a number of biomedical application-oriented parser evaluations. Evaluation of non-native GR/SD parsers on resources annotated using these representations requires conversion, and although not all such evaluations have included estimates of the quality of conversion (e.g. Preiss, 2003; Kaplan et al., 2004), a number of studies assessing the difficulty of various conversions have recently been published. Their results are summarized in Table 4.2. While a number of factors relating to e.g. conversion methodology prevent strong conclusions from being drawn, these results are consistent with the hypothesis that the GR scheme represents a difficult conversion target (for discussion, see Clark and Curran, 2007; Sagae et al., 2008a), while the creation of accurate conversions into SD is feasible. This result may be explained in part by the different emphasis of the representations on deep vs. surface-oriented structure. If similar results hold also in future evaluations, they present an argument in favor of SD for formalism-independent evaluation.

Second, recent studies evaluating parsers on biomedical text using SD and either the GTB corpus (Clegg and Shepherd, 2007; Sagae et al., 2008a) or BioInfer (Haverinen et al., 2008) allow the picture of parser performance and adaptation methods painted in Papers I–III to be filled in further. As discussed above, Clegg and Shepherd evaluated nine variants of five statistical treebank parsers, and Sagae et al. evaluated the HPSG parser Enju, the Charniak-Johnson reranking parser (Charniak and Johnson, 2005) and the Charniak parser (Charniak, 2000). In our recent study using the methodology presented in Paper III to create a Pro3Gres→SD conversion (Haverinen

et al., 2008), we evaluated the Pro3Gres parser (Schneider, 2007), which has been developed with particular attention to the challenges of parsing biomedical domain text, comparing its performance to that of the Charniak-Lease parser. Sagae et al. performed also the important experiment of retraining the statistical parsers (with the exception of the Charniak-Johnson reranker) on GTB. Although not a lightweight adaptation approach in the sense studied in Paper II in requiring an annotated domain treebank, this experiment provides important complementary information regarding domain adaptation. The largely comparable results of these studies suggest that while there is not much difference among the best-performing parsers, the Charniak-Lease parser performs the best among those evaluated "out of the box,"—without retraining the parser—as parsers are most commonly applied in domain studies. Of those evaluated after training on a domain treebank, the Enju and Charniak-Johnson parsers jointly achieve the best results. A recent task-oriented evaluation of eight parsers (not including Charniak-Lease) by Miyao et al. (2008) suggests roughly similar conclusions regarding the merits of the parsers for PPI extraction, with Charniak-Johnson performing best without retraining. When trained on GENIA, the best-performing parsers were Enju with the adaptations of Hara et al. (2007) and the native dependency parser of Sagae and Tsujii (2007).

Third, these results allow a note on the effect of training on a domain treebank: both studies report performance for the Charniak parser using largely the same setup, SD conversion ("collapsed" output) and the GTB corpus, although using different subsets of GTB. Clegg and Shepherd (2007) report 77.0% F-measure without retraining and Sagae et al. (2008a) 81.2% when the parser is trained on GTB. These results suggest an 18% relative reduction in error from training on a domain treebank. In view of the efforts required for treebank annotation, this improvement appears somewhat modest, bringing into question whether treebank annotation is an effective focus of efforts for adaptation to new domains (see also Sagae et al., 2008b). However, this estimate should be taken cautiously, as subtle differences in evaluation strategy can cause substantial differences in results (see e.g. the discussion of collapsing in Paper III, Section 5.3).

Fourth, as Sagae et al. (2008a) evaluate the same parsers both trained and tested on PTB and trained and tested on GTB, the results provide an estimate of the difficulty of parsing biomedical domain text. The best reported SD result for PTB is 88.4% and for GTB 82.0% F-measure, for an estimated 55% increase in error from general English to biomedical text, even including training on a domain treebank. While the differing sizes of the two treebanks are a confounding factor, this result is in line with the rough estimate provided in Paper I and suggests further that additional efforts are still required for biomedical domain parsers to reach the level of accuracy achieved in parsing news domain English.

In conclusion, while Papers I–III and other recent studies have shed light into the issues of parser performance on biomedical text, domain adaptation methods, and the feasibility of unification under shared syntactic representations, several open issues remain. Compared to the relatively steady state in parser evaluation methodology in the decade following the introduction of the PARSEVAL measures, parser evaluation is currently in something of a state of flux, with movement toward formalism-independent dependency evaluation but more variance in results. The results of Paper III support the value of the SD scheme as a unifying representation, in particular for application-oriented studies, and the conversion and annotation work in the study has created the first native SD gold standard; this data is distributed as part of the BioInfer corpus. This resource may provide a valuable reference point for standardizing domain evaluation and development efforts.

Chapter 5

BioInfer corpus

The work described in Paper IV, the design and annotation of the BioInfer (Bio INFormation Extraction Resource) corpus, was the largest single subproject of those described in this thesis. When the first efforts to produce an annotated corpus started in our group in 2001, the motivation for manual annotation was simple: there were only few publicly available biomedical corpora, and none that could fully support the development and evaluation of a text mining system using full dependency parsing: for this purpose, a corpus must minimally contain annotation for named entities and their relationships, and syntactic annotation is necessary for evaluating and developing parsers for the task. Combining these annotations for a single set of sentences further allows the interplay of parsing, named entity recognition and relation extraction to be studied in detail. These three classes of annotation, entities, relationships, and syntax, form the core of the BioInfer corpus annotation.

The following sections briefly introduce BioInfer and Paper IV. Due to the complex nature of parts of the corpus annotation, the short description in this chapter is not an attempt to present a detailed definition of BioInfer annotation, but rather to motivate some of the central design decisions and present an overview of the corpus. The main contribution of this work, the BioInfer corpus itself, is freely available at the corpus web page, `http://www.it.utu.fi/BioInfer`.

5.1 Named entities

BioInfer entity annotation is built around named entities of the protein, gene and RNA types. Only specific, established names such as *actin* are annotated, not underspecified references such as *a 50 kDa protein*. We do not, by design, annotate nesting (embedding) in names: for example, *MAP kinase* is annotated as a simple name with no separate annotation for *MAP*. This decision stems from the view that the names are, once established, typically no longer either used or understood as descriptive. The alternative of annotating names with full nesting together with the general aim of the corpus annotation to capture all stated relationships would require, at one extreme, each mention of e.g. *mitogen-activated protein kinase kinase kinase (MAP-KKK)* to be annotated as stating four relationships, if not more (Rzhetsky et al., 2004); this choice would not reflect how biologists understand simple name mentions. The decision not to annotate name internal structure agrees with the strategy taken in PennBioIE annotation (Kulick et al., 2004) but differs from the approach taken, for example, in the annotation of the AIMed corpus (Bunescu et al., 2005).

The most commonly applied annotation scheme for named entities only allows markup of continuous spans of text. For MUC annotation, for example, the text *North and South America* would be annotated as containing the two locations *South America* and *North* (Sundheim and Chinchor, 1995). This strategy has been found inadequate for accurate annotation of biomedical text by several groups working in the domain. The BioInfer annotation captures names precisely even in constructs involving, for example, coordination with elision of a head word (*alpha and beta catenin* → *alpha catenin*, *beta catenin*) or breaking syntactic token boundaries (*Arp2/3* → *Arp2*, *Arp3*). Similarly detailed entity annotation has been produced also for the GENIA (Ohta et al., 2002) and PennBioIE (Kulick et al., 2004) corpora, although applying quite different annotation schemes.

5.2 Entity relationships

The prevailing approach to relationship annotation in corpora for biomedical text mining is simply to specify which pairs of annotated named entities are connected by some form of relationship. This information may be augmented by specifying, e.g. directed pairs or the type of the relationship, as discussed further in Chapter 6. While pairwise annotation is a reflection of mainstream practice in current biomedical IE, it fails in many cases to accurately capture the information stated in text, even when typed, directed pairs are annotated.

Consider the following examples of possible typed pairwise annotation. This annotation is sufficient to represent simple statements:

e_1 interacts with e_2 → INTERACT(e_1,e_2)

e_2 is phosphorylated by e_1 → PHOSPHORYLATE(e_1,e_2)

some approximation is necessary, but likely to be acceptable in some more complex cases:

abundance of e_2 is affected by e_1 expression → AFFECT(e_1,e_2)

e_1 inhibits e_2 activity → INHIBIT(e_1,e_2)

meaningful approximation is more difficult for statements that include additional properties or processes involving the named entities:

e_1 promotes e_2 polymerization → ? (e_1,e_2)

e_1 inhibits phosphorylation of e_2 → ? (e_1,e_2)

here, while the creation of "complex types" for annotation relations, e.g. INHIBIT-PHOSPHORYLATION, can serve as a stopgap solution preserving the details of the relationship (see Heimonen et al., 2008), this approach gains simplicity in the pairwise scheme only at the cost of a potentially open-ended inventory of relationship types. Finally, pairwise annotation cannot capture complex relationships involving more than two entities. As an example, in sentences such as *activation of* e_1 *by* e_2 *prevents the phosphorylation of* e_3 *by* e_4 there is no single, obviously correct decision on whether and how to annotate the pairs (e_1,e_3), (e_1,e_4), (e_2,e_3) and (e_2,e_4), and there are apparent inconsistencies between corpora in how complex relationships such as this are annotated.

An early unpublished version of the BioInfer corpus relationship annotation was produced using a basic annotation scheme allowing only pairwise, untyped interactions between named entities. However, later attempts to assign accurate types to the relationships and to make the initial annotation more consistent proved problematic. Annotators with a background in biology, in particular, refused to accept many of the considered approximations as meaningful. Additionally, attempts to formulate consistent guidelines for pairwise annotation of complex relationships were frustrated by repeated occurrences of exceptional cases that fell outside the scope of what could be naturally captured in the annotation scheme or anticipated in the guidelines. We ultimately reached the conclusion that the pairwise annotation scheme was not sufficiently expressive to represent the statements found in the corpus.

The current BioInfer relationship annotation is based on a scheme that allows complex, structured relationships to be explicitly annotated. The

annotation scheme captures the above example sentence *activation of* e_1 *by* e_2 *prevents the phosphorylation of* e_3 *by* e_4 as

PREVENT(ACTIVATE(e_2,e_1), PHOSPHORYLATE(e_4,e_3)).

Interestingly, we found that creating and following consistent rules for producing this more complex annotation was more straightforward than for the simple pairwise annotation scheme. The BioInfer annotation manual (Ginter et al., 2007) describes the annotation rules for named entities and relationships in detail.

In addition to annotation for named entities, the BioInfer scheme includes annotation for several other types of physical entities as well as abstract process and property entity types. The entities participating in relationships are marked up for their full internal structure, with named entities as their atomic innermost core. For example, e_1 *prevents the initiation of* e_2 *polymerization* would be annotated as containing not only the named entities e_1 and e_2 but also the process entities e_2 *polymerization* and *initiation of* e_2 *polymerization*, with the PREVENT relationship annotated as holding between e_1 and the *initiation* process. All annotated entities and relationships in the BioInfer corpus are assigned the most specific applicable types from ontologies, which are further designed to make it possible to formulate systematic rules for decomposing such complex relationships for applications that require entity pair annotation. These ontologies are briefly described next.

5.3 Ontologies

BioInfer defines two ontologies[1]: an entity type ontology that incorporates the established GENIA ontology of physical types (Ohta et al., 2002) and extends it with abstract types (processes and properties), and a relationship type ontology that has been created specifically for the BioInfer annotation needs. Both ontologies are hierarchically structured, with *Entity* and *Relationship* as the most general types and, for example, *Gene* and DOWN-REGULATE among instances of the most specific types. The backbone of the BioInfer relationship type ontology, showing classes of relationship types, is shown in Figure 5.1, and a simplified fragment of the entity type ontology in Figure 5.2.

[1]We understand "ontology" generally in the sense of Mitkov (2003, page 750): *"An inventory of the objects or processes in a domain, together with the specification of some or all of the relations that hold among them, generally arranged as a hierarchy"*. This usage is common in the domain: Hersh et al. (2004) describe the Gene Ontology as *"not an ontology is the purists' sense,"* but a hierarchically organized controlled vocabulary. This description applies also to the BioInfer ontologies, although in terms of size and scope they are far from GO.

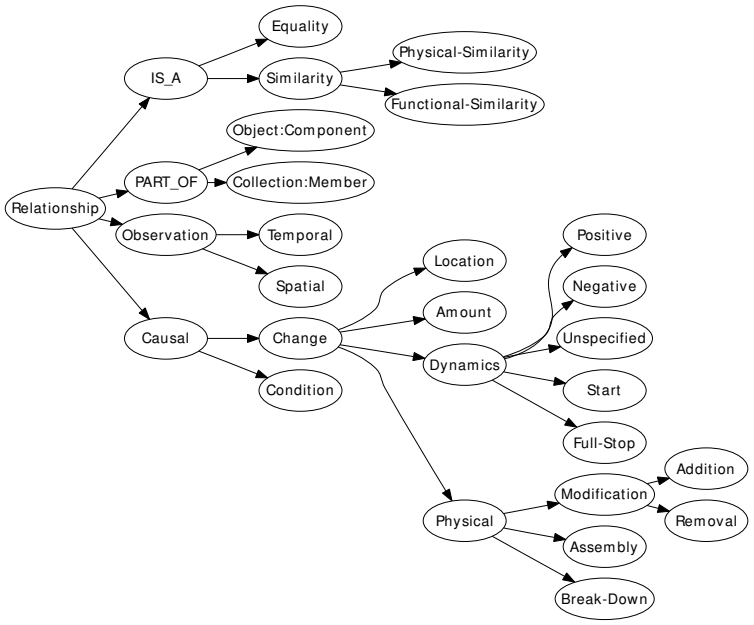

Figure 5.1: *The backbone of the BioInfer relationship type ontology.*

The ontologies are organized in part around the concept of entity state, understood as consisting of *amount*, *location*, *dynamics* and *physical* properties. In addition to serving as an organizing principle, entity state further provides a connection between the ontologies: for example, the entity e_1 *phosphorylation* is annotated as a process of the type PHOSPHORYLATION, which provides a connection to the PHOSPHORYLATE relationship type assigned to statements such as e_1 *phosphorylates* e_2. It should be noted that the inclusion of processes in the entity type ontology can be seen as somewhat nonstandard. Under the philosophical distinction between *continuants* (things that endure through time, undergoing change) and *occurrents* (which occur and unfold in time), processes fall perhaps more naturally together with occurrents (i.e. in the relationship type ontology for BioInfer) than with continuants such as *Gene*. This view has been recently advocated in biomedical ontology construction by Smith et al. (2005) and adopted for the GENIA event ontology (Kim et al., 2008a). We note that the classification of processes together with continuants, for BioInfer motivated by annotation considerations and the observation that processes, like continuants, are often stated to undergo change, is not entirely without proponents (Galton, 2006). Nevertheless, the connection between process entities and their corresponding relationships in the BioInfer ontologies allows these entities to be interpreted as (underspecified) relationships if necessary, a possibility we have pursued in recent work (Heimonen et al., 2008).

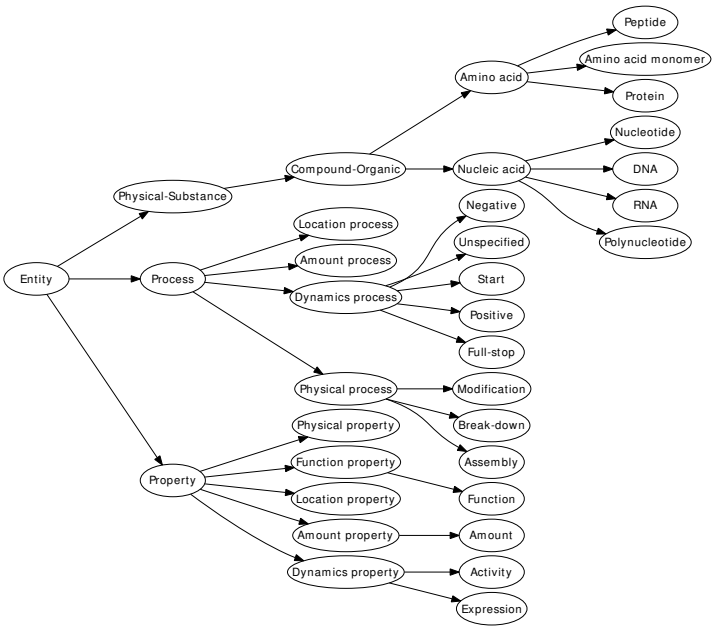

Figure 5.2: *A simplified fragment of the BioInfer entity type ontology. A part of the GENIA ontology of physical types appears as the* Physical-Substance *branch. The* Process *branch mirrors the* Causal-Change *branch of the BioInfer relationship type ontology.*

5.4 Other contributions

The third major class of annotation in the BioInfer corpus is the syntactic annotation, originally created in the LG scheme (see Papers I and II) and, after the publication of the first release of the corpus, extended to include annotation also in the SD scheme (Paper III). Comparisons of these schemes in found Chapter 4 of this thesis and in Paper III. A specific contribution toward the syntactic annotation of the BioInfer corpus that is only described in Paper IV is the introduction of a reliable heuristic method for assigning dependency types to a corpus annotated with untyped dependencies, described in Paper IV, Section *Dependency types*.

In addition to entity, relationship, and syntactic annotation, the BioInfer corpus additionally contains partial coreference annotation, annotated using the relationship formulas. The relationship annotation mechanism is also employed to capture aliasing and abbreviations and other static, non-event relationships such as protein family membership or part-whole relationships, which are rarely annotated in biomedical domain corpora. BioInfer includes annotation also for explicitly negated statements of relationships between entities: for example, e_1 *does not bind* e_2 is annotated as NOT(BIND(e_1,e_2)).

Entities		Relationships		Dependencies	
Total	6349	Total	2662	Total	28139
named entities	72%	*causal*	55%	coverage	94%
		is_a	14%		
		part_of	22%		
		observation	5%		
		other	4%		

Table 5.1: Selected BioInfer corpus statistics [From Paper IV]. Dependency annotation coverage is the ratio of dependencies to non-punctuation tokens.

More detailed descriptions and examples of these annotations are found in Appendix I of Paper IV. Finally, relationships, as well as entities, are annotated to mark the specific words expressing them in the text of the corpus sentences. While the specific occurrences of named entities are almost universally marked in corpora that annotate them, BioInfer extends this principle to include relationships, so that, for example, e_1 *is a cofactor for* e_2 is annotated as BIND(e_1,e_2) with the word *cofactor* marked as expressing the BIND relationship.

The complex, multifaceted nature of the BioInfer annotation is reflected to some extent in the format of the corpus data. To preserve the original sentence structure as well as to allow for different divisions of the text for different layers of the annotation, the corpus follows the standoff annotation principle, where the original text is unmodified and different annotations are marked with character offset references into the text. To allow the processing of the corpus with standard tools, it is distributed in XML format, a standard for structured data. Finally, to aid users of the corpus to access different aspects of the corpus data and to view the annotations together, we provide software tools for extracting annotations in simplified formats and visualizing the corpus data, openly distributed with full source along with the corpus.

5.5 Discussion and conclusions

The initial release of the BioInfer corpus contained annotation for 1100 sentences; some key statistics of this annotation are given in Table 5.1. While larger corpora containing some of the BioInfer annotation types, for example named entity annotation, had been available for some time, BioInfer was the first corpus in the domain to combine entity, relationship and syntactic annotation in a corpus of this size. Additionally, BioInfer was the first available domain corpus to provide annotation for complex, structured relationships. In combining detailed relationship and entity annotation with

relationship typing, the BioInfer corpus took a small step toward annotation that connects text to a *knowledge representation* in the sense of a computable model that can support inference. As a partial validation of this capacity, we have recently implemented inference rules that aim to deduce the underlying physical and regulatory relationships between named entities on the basis of the complex BioInfer relationship annotation (Heimonen et al., 2008).

Several related corpora published prior to BioInfer are discussed in Paper V and Chapter 6, but this excludes a key piece of related work, the recently published version of the GENIA corpus that includes event annotation (Kim et al., 2008a). This annotation largely corresponds in its expressiveness to the BioInfer relationship annotation, including also structures where entities affect events involving other entities. A number of other aspects are shared by the BioInfer and GENIA efforts, such as the annotation of spans of text stating relationships—termed *text binding* in BioInfer and *text-bound annotation* in GENIA—and the use of an annotation scheme that can be applied by biologists without reference to particular linguistic theories. As these two resources have been developed in parallel but independently of each other, these convergences are encouraging and can be seen as validating many of the design choices made. The GENIA event annotation covers a set of sentences roughly an order of magnitude larger than BioInfer, though it is somewhat less comprehensive in the scope of the annotated relationships: GENIA focuses on "dynamic" relations, excluding static relationships such as *part_of* and *is_a*. Its scope thus roughly corresponding to the types in the *Causal-Change* branch of the BioInfer relationship ontology, by which it can be estimated to cover 55% of the relationships annotated in BioInfer (see Table 5.1).

The annotations of the BioInfer and GENIA event corpora present an opportunity to address a number of challenges in domain IE. While carefully crafted hand-written systems capable of extracting complex relationships between biomedical entities have been presented (Friedman et al., 2001; Hunter et al., 2008), learning to reliably extract such relationships remains an open problem. In our recent work we have begun to pursue a semantic network-based approach building on a dependency parse representation of syntactic structure to address this challenge (Björne et al., 2008); we expect that the BioInfer and GENIA corpora will provide critically important resources to the development and evaluation of this and other approaches to advanced biomedical IE.

Chapter 6

Protein-protein interaction extraction

The extraction of protein-protein interactions (PPIs) has been the most widely studied IE task in biomedical text mining for almost a decade now, and a great number of methods have been proposed and several corpora have been made publicly available. However, the field still lacks accepted standards for annotation, evaluation methodology, or comparison of different tools and techniques. As a result, each corpus is annotated using different, often either unwritten or unpublished annotation guidelines and published in a different format. The reported performance results for methods developed and evaluated on different resources are largely incomparable, and it is difficult, if not impossible, to reliably judge the relative merits or effectiveness of the various proposed PPI extraction methods. This situation provided the impetus for the study presented in Paper V.

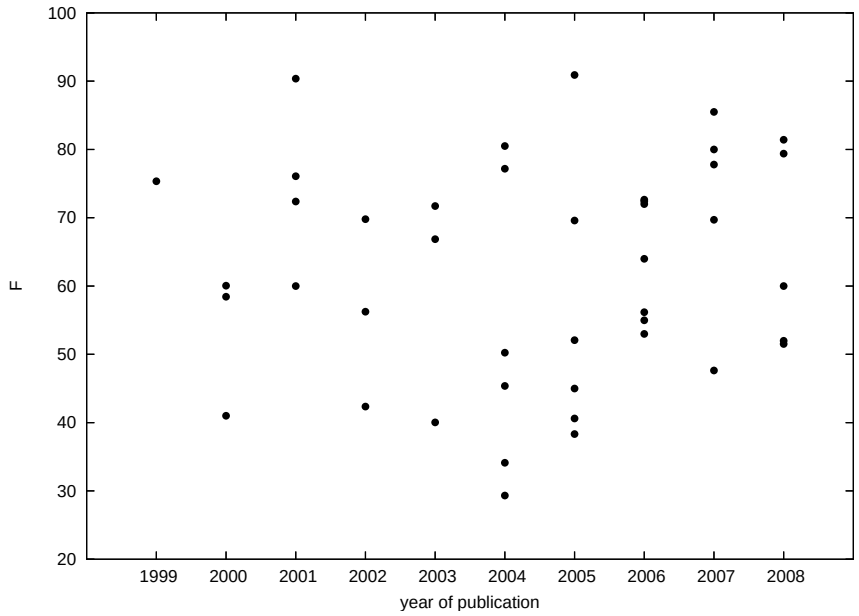

Figure 6.1: *Plot of reported protein-protein interaction extraction method performance results.*

6.1 Extraction method evaluation

The lack of widely shared standard resources and evaluation methodology can be vividly seen in the results reported for PPI extraction system performance over the last ten years. Figure 6.1 illustrates the reported results found in a survey of studies published outside of shared task evaluations such as LLL and BioCreative, showing the best result included in each study and omitting those for which F-measure could not be calculated (data in Table 6.2 in the Appendix).[1] It is immediately obvious that, contrary to what one would hope, there is no clear trend toward better performance over the years to be found in these results. One of the highest results included is reported for the relatively early system of Ono et al. (2001), and in the recent large-scale PPI extraction evaluation in BioCreative II (Krallinger et al., 2007) the best-performing system (Hunter et al., 2008) achieved only a 29% F-measure, performance worse than almost all of the results reported in the studies in Table 6.2.

This broad dispersal of results represents a failure of consistent evaluation, the magnitude of which is particularly apparent when the results are

[1]It should be noted that this is not proposed to constitute an exhaustive survey, and that this compilation is not intended to imply that the authors of the studies claim that these results are directly comparable—even though comparability is implied by frequently-seen statements such as "state-of-the-art performance."

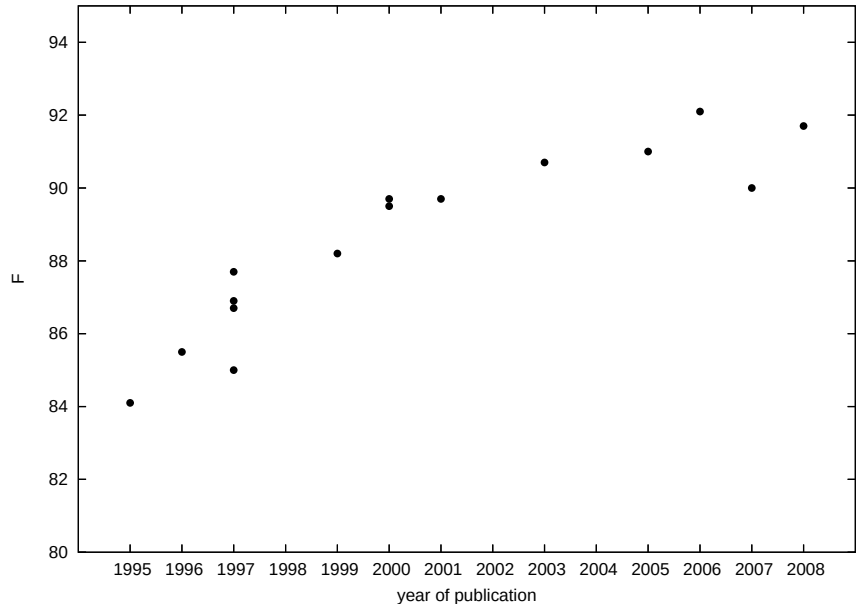

Figure 6.2: *Plot of selected parsing performance results.*

compared to those reported for a field that has shown consistent improvement over a similar period, such as parsing efforts focused on the Penn Wall Street Journal treebank, where a decade of study has roughly halved the error rate at the task. A selection of results from these studies are shown in Figure 6.2 (data in Table 6.3 in the Appendix). Pallett et al. (2000) provide a striking illustration of similar progress in speech recognition.

The conflicting results reported for different PPI extraction systems were noted relatively early: Park (2001) notes *"We found curious performance differences between our system and that of [Ono et al. (2001)]"* regarding the better reported recall and over 40 percentage unit advantage in precision for the system that, contrary to that proposed in the study, attempted no linguistic processing. In their well-heeded call for shared task evaluations, Hirschman et al. (2002b) noted simply that *"it is unclear how to compare the different approaches"*. While shared tasks such as LLL and BioCreative have since provided a partial response to the problem of comparability, it would be a considerable loss to retreat to the position that other results simply cannot be compared; nearly 50 such studies are referenced in Table 6.2, and many more are likely to follow each year. One alternative is to attempt to identify the sources of the variation in estimated performance results and quantify the magnitude of their effect on measurements, as a step toward controlling the undesired variance. This is the approach taken in Paper V.

6.2 Comparative corpus evaluation

Of the many possible explanations for the substantial divergence in reported results for similar methods, the corpus on which the methods are evaluated is perhaps the most obvious candidate. In the study reported in Paper V, we performed the first comparative evaluation of PPI corpora. We gathered publicly available, manually annotated PPI corpora that contained a sufficient level of annotation to be used for the training and evaluation of methods for extracting PPIs between specific named proteins (Paper V, Section *Corpora*). We studied the annotations found in the corpora, converted them into a shared common format, and performed quantitative and qualitative analyses to characterize their differences.

To evaluate the effect of the choice of corpus on estimated performance, we reversed the typical evaluation setting: instead of using a corpus as a benchmark to evaluate a PPI extraction methods, we used PPI extraction methods as benchmarks to evaluate corpora. To select an extraction method from among the large number proposed, we considered a number of factors. First, for the results to be relevant, a relatively recent method with state-of-the-art performance[2] was required.

Second, an implementation of the method was needed to be either publicly available or sufficiently straightforward so that a faithful reimplementation could be performed. Further, the method was required to cover a range of interaction types in order to be applicable to several corpora. Finally, methods involving hand-written rules were preferred to machine-learning methods, as the performance of the former is not affected by corpus size—for machine-learning methods, the effect of differing amounts of training data would have to be controlled, and the stability of results on small corpora might be limited. We settled on the RelEx method of Fundel et al. (2007), which boasts these and a number of other positive attributes.

RelEx makes use of full parsing, and its extraction rules are based on a list of interaction-expressing words and paths connecting proteins to such words in an SD dependency representation of sentence structure. The three core extraction rules are illustrated in Figure 6.3, detailed descriptions are given in Paper V and the original study of Fundel et al. (2007). We reimplemented the RelEx method with gracious help from Fundel, who provided both advice and data to help in the development.

[2]This involves a circular argument that is difficult to avoid. To increase the chance of selecting a method with good "true" performance, we considered evaluation on more than one corpus a merit.

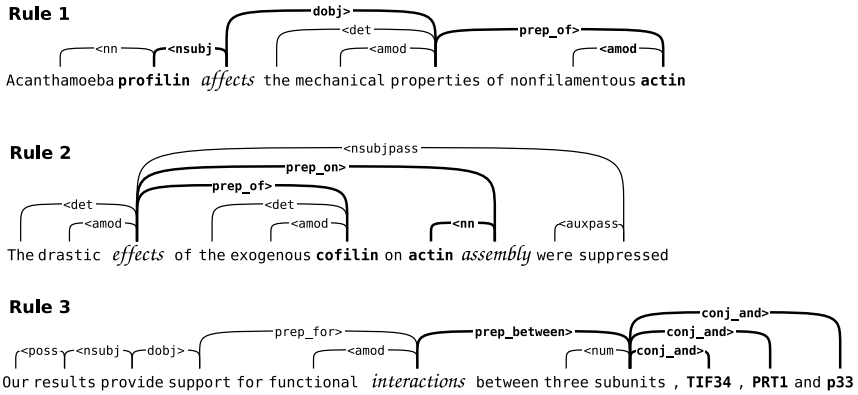

Figure 6.3: *Illustration of RelEx interaction extraction rules. Protein names are shown in bold, possible interaction words in italics, and interaction-expressing paths with thick lines. [Figure from Paper V]*

6.3 Results

We selected five corpora for the analysis: AIMed (Bunescu et al., 2005), BioInfer (Paper IV), HPRD50 (Fundel et al., 2007), IEPA (Ding et al., 2002) and LLL (Nédellec, 2005). We first studied the corpora to identify the "greatest common factor" level of information provided regarding PPIs in these corpora: undirected, untyped interactions with no information regarding the words expressing the interaction, no complex structure, and no annotation for negations or the certainty of PPI statements (Table 6.1[3]). All corpora were then converted by custom-written software into a shared format capable of capturing this information. We then evaluated the performance of the RelEx method on each corpus, applying also a simple baseline method that assigns an interaction to all proteins co-occurring in a sentence (i.e. *all-true*). The results of this evaluation are illustrated in Figure 6.4.

While differences in measured performance between different corpora was expected, their magnitude is striking. The performance of RelEx differs on average by almost 20% between pairs of corpora, while the average difference between the trivial co-occurrence baseline and the state-of-the-art RelEx method is less than 15%. This, perhaps the single most important finding of this study, implies that unless the effect of the choice of corpus on measured performance is controlled, the direct comparison of performance results for methods evaluated on different resources is essentially meaningless.

[3]There is an error in the table from which this data has been taken, Table I, in the preprint of Paper V included in this thesis. The row *PPI types* in the table should read *no, 68 types (ontology), no, no, 3 types*—BioInfer, not HPRD50, uses an ontology of PPI types. This error has apparently been introduced in the BMC Bioinformatics PDF production process, as it does not occur in the HTML version of the paper.

		AIMed	BioInfer	HPRD50	IEPA	LLL
	size	1955	1100	145	486	77
PPI	*types*	no	yes	no	no	yes
	binding	no	yes	no	yes	no
	directed	no	yes	no	yes	yes
	complex	no	yes	no	no	no
	negative	no	yes	no	no	no
	certainty	no	no	yes	no	no

Table 6.1: Corpus size in sentences and characteristics of the PPI annotations in the analysed corpora:
types: explicit indication of the type of the annotated interactions
binding: identification of the text spans that state the interactions
directed: specification of the directionality of the interaction
complex: annotation includes nested or n-ary (for $n > 2$) interactions
negative: annotation of negative interactions
certainty: annotation of levels of certainty of interactions
[From Paper V]

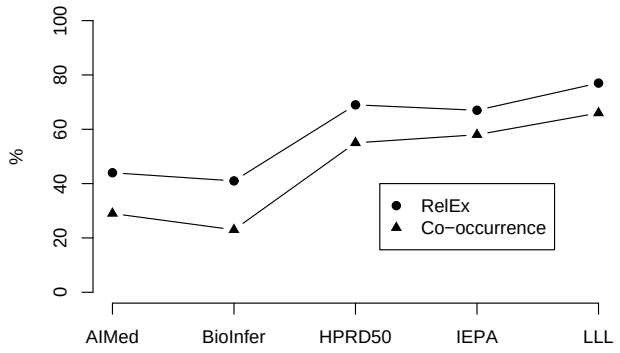

Figure 6.4: *RelEx and Co-occurrence extraction performance on different corpora. [Figure from Paper V]*

In Paper V, we proceeded to study the characteristics of the corpora in detail to determine the major sources of this variance. We found that there is considerable variance in the ratio of interactions to proteins pairs annotated in the different corpora, a factor that we estimated to explain almost half of the observed difference in performance between corpora. This ratio is important because the F-measure metric that is almost exclusively applied

to evaluate PPI extraction performance is sensitive to the distribution of positive cases (interactions) to negative cases (pairs of proteins that do not interact). The wide variance in the distribution of positive and negative cases between corpora indicates that the F-measure is a poor choice of metric for the purpose of comparing performance between corpora. The use of a distribution-independent measure such as AUC (area under the receiver operating characteristic curve; see e.g. Hanley and McNeil, 1982), considered for biomedical domain evaluations by Hirschman et al. (2002b), could address this issue and considerably improve comparability, as we demonstrate in (Airola et al., 2008). The differences in the distribution of positive and negative examples between the corpora may reflect different assumptions on the part of the corpus annotators on how the PPI extraction task should be approached, in particular regarding on how effectively irrelevant documents can be filtered out and whether entire documents or only separately filtered relevant sentences are to be processed by the PPI extraction component.

In the study reported in Paper V we further performed a qualitative analysis of the corpora, considering characteristics of the PPI statements such as the type of the interactions, whether they are direct or indirect, and whether they are explicitly and definitely stated. This analysis tentatively suggested that the explicitness of the PPI statements might affect PPI extraction performance, but no simple connection of the other factors with measured performance was observed, and no statistically significant effects were found in this analysis.

6.4 Discussion and conclusions

The work reported in Paper V provided the first comparative evaluation of available PPI corpora, establishing the magnitude of the effect of the choice of corpus on the evaluated results of PPI extraction methods. A detailed evaluation of possible sources for these differences indicated that the distribution of positive and negative instances of annotated PPIs varies notably between corpora and contributes almost half of the variance observed in measured performance. This further calls into question whether the ubiquitous F-measure metric is appropriate for comparative PPI extraction method evaluation.

One possible response to the findings of Paper V would be to focus on results evaluated on each corpus separately, under the assumption that these will be comparable. A possible candidate for a corpus that has found sufficiently large use to serve as the basis of comparison is AIMed, which has been applied in numerous domain studies (including Bunescu et al., 2005; Bunescu and Mooney, 2005; Ramani et al., 2005; Yakushiji et al., 2005; Bunescu and Mooney, 2006; Giuliano et al., 2006; Katrenko and Adriaans,

2006; Mitsumori et al., 2006; Yakushiji et al., 2006; Erkan et al., 2007; Sætre et al., 2007; Airola et al., 2008; Van Landeghem et al., 2008; Miyao et al., 2008) and may be seen as an emerging *de facto* standard in particular for the evaluation of machine learning methods for biomedical relation extraction. However, Sætre et al. (2007) recently demonstrated that differences in evaluation protocol—specifically, in how cross-validation is performed—can make a difference of almost 20 percentage units in results measured on the AIMed corpus. Even if the choice of corpus and the evaluation protocol are held constant, different preprocessing of the corpus can lead to an almost 30% difference in the number of negative cases generated from AIMed (Sætre et al., 2007, Table 3). This will in turn be reflected in F-measure evaluation results, as we recently demonstrated in a study focusing on AIMed results for establishing comparative performance of a PPI extraction method (Airola et al., 2008). Finally, differences in the definition of correctness criteria for extraction (see e.g. Giuliano et al., 2006) and, to a lesser extent, in parameter selection for machine learning methods can introduce bias into evaluation. These issues indicate that biomedical relation extraction is in need of an evaluation standard of the type that the WSJ section of the Penn Treebank and the PARSEVAL measures—whatever their other faults—have provided for statistical parsing. Building on the work reported in Paper V, we recently proposed such a standard (Pyysalo et al., 2008).

As part of the work to perform the evaluation in Paper V, we created custom software for converting each of the AIMed, BioInfer, HPRD50, IEPA and LLL corpora into a common, shared representation, thus unifying the various annotations and allowing these corpora to be used together or combined into a large, multi-domain corpus. This unification considerably increases the amount and variety of data easily available for training and evaluating PPI extraction methods. It also provides an opportunity to test new methods on multiple corpora with little additional effort. Such broader testing could, to an extent, address the current difficulty of establishing the comparative performance of PPI extraction methods and serve as a step toward more meaningful comparisons. It is encouraging to note that the unified corpora have already generated a measure of interest in the community and that some studies making use of this recent contribution have been carried out.

Appendix

This appendix contains the results of a survey of protein-protein interaction extraction methods and a selection of statistical parsing results. See Section 6.1 for context.

	precision	recall
(Sekimizu et al., 1998)	73	-
(Rindflesch et al., 1999)	79	72
(Proux et al., 2000)	87	44
(Thomas et al., 2000)	70	29
(Rindflesch et al., 2000a)	73	51
(Friedman et al., 2001)	96	63
(Ono et al., 2001)	94	87
(Park et al., 2001)	80	48
(Stephens et al., 2001)	89	61
(Yakushiji et al., 2001)	-	49
(Blaschke and Valencia, 2002)	45	40
(Leroy and Chen, 2002)	70	47
(Pustejovsky et al., 2002)	90	57
(Palakal et al., 2002)	81	-
(Ding et al., 2003)	87	61
(Koike et al., 2003)	87	26
(Leroy et al., 2003)	90	-
(Temkin and Gilder, 2003)	70	64
(Corney et al., 2004)	55	20
(Daraselia et al., 2004)	91	21
(Huang et al., 2004)	81	80
(Karopka et al., 2004)	93	30
(McDonald et al., 2004)	89	35
(Rzhetsky et al., 2004)	95	65
(Ahmed et al., 2005)	66	27
(Bunescu et al., 2005)	45	45
(Hao et al., 2005)	81	61
(Plake et al., 2005)	60	46
(Xiao et al., 2005)	88	94
(Yakushiji et al., 2005)	37	45
(Bunescu and Mooney, 2006)	53	53
(Giuliano et al., 2006)	65	63
(Jang et al., 2006)	81	43
(Katrenko and Adriaans, 2006)	75	70

Continued on Next Page...

Table 6.2 – Continued

	precision	recall
(Mitsumori et al., 2006)	56	54
(Rinaldi et al., 2006)	90	60
(Yakushiji et al., 2006)	64	84
(Zhou et al., 2006)	73	48
(Erkan et al., 2007)	86	85
(Fundel et al., 2007)	80	80
(Sætre et al., 2007)	78	63
(Sun et al., 2007)	82	74
(Yang et al., 2007)	55	42
(Alex et al., 2008)	51	53
(Airola et al., 2008)	73	87
(Clegg, 2008)	57	47
(Kim et al., 2008c)	73	83
(Van Landeghem et al., 2008)	79	84
(Miyao et al., 2008)	55	66

Table 6.2: Reported protein-protein interaction extraction method performance results 1998–2008. Studies reporting neither precision nor recall not included. For studies reporting results for multiple methods or datasets the best result is given, for those reporting several samples or an estimated performance range, the average result or midpoint is given, and for those reporting precision-recall curves, an estimated break-even point is given.

	P	R	F
(Magerman, 1995)	84.0	84.3	84.1
(Collins, 1996)	85.3	85.7	85.5
(Charniak, 1997)	86.7	86.6	86.7
(Collins, 1997)	87.5	88.1	87.7
(Goodman, 1997)	84.8	85.3	85.0
(Ratnaparkhi, 1997)	86.3	87.5	86.9
(Collins, 1999)	88.1	88.3	88.2
(Charniak, 2000)	89.6	89.5	89.5
(Collins, 2000)	89.6	89.9	89.7
(Bod, 2001)	89.7	89.7	89.7
(Bod, 2003)	90.8	90.7	90.7
(Charniak and Johnson, 2005)	-	-	91.0
(McClosky et al., 2006)	-	-	92.1
(Petrov and Klein, 2007)	89.9	90.2	90.0
(Huang, 2008)	-	-	91.7

Table 6.3: Selected parsing results on Penn Wall Street Journal treebank, labeled (P)recision, (R)ecall and F measures. For each study, performance shown for the largest of the ≤ 40 *words*, ≤ 100 *words* and *complete* subsets of the test corpus for which results were reported.

Chapter 7

Conclusions

The preceding chapters described five studies exploring aspects of biomedical text mining with a particular focus on the use and value of full dependency parsing to protein-protein interaction extraction.

Papers I and III addressed the issue of identifying the best tools and methods for performing dependency parsing of biomedical text. While dependency parsers had been widely used in the domain, there was little information regarding their performance when this research was started, and our initial study on this task provided the first detailed evaluation of a dependency parser on a fully annotated domain corpus. Several studies on domain parser performance evaluation have been published since, including many whose results can be directly compared due to the use of a common representation of dependency structure. The work presented in Paper III provided support for the feasibility and value of such unification, demonstrating that highly accurate conversions between different dependency schemes can be created. While there are still several open questions regarding the challenges of parsing biomedical text, I believe that, in addition to establishing the performance of two widely-applied parsers, these studies have contributed toward clarifying the broader issues and provided support for a common, application-oriented evaluation strategy.

The analysis of parser failures in Paper I and the comparative analysis of lexical adaptation methods in Paper II identified many of the challenges that general English parsers face in biomedical text and established the relative merits of a number of approaches to resolving issues related to domain vocabulary. This work was done specifically in the context of a parser with a broad-coverage hand-written grammar of general English, and has been complemented by a number of studies that have studied the lexical adaptation of a statistical treebank parser, the effect of retraining statistical parsers on a biomedical domain corpus, and the effectiveness of reranking methods for identifying better parses among the ambiguous alternatives generated by

parsers. As part of the work reported in Paper III, we further demonstrated that a surface-oriented dependency representation can be accurately converted into a more semantically oriented scheme, increasing the value of a popular parser for further applications without loss of parse quality. In addition to identifying and addressing problems in parsing domain text, these studies have contributed better tools for biomedical text mining researchers, as all resulting tools have been made freely available.

Evaluation techniques were considered and the challenges of establishing comparability addressed in particular in Papers I and III in the context of parsing and in paper V for protein-protein interaction extraction. In Paper I two parsers using very different representations were evaluated intrinsically, each using largely its own scheme, but with modifications and simplifications that balanced the number of structural constraints that each parser needs to meet to assure comparability. Paper I also proposed the use of a task-oriented metric to provide a measure of expected performance at a domain information extraction task. Papers III and V involved a different approach to evaluation: unification under a shared representation. The methodology applied in Paper III was shown to make it possible to create a conversion between different syntactic representations that has one of the highest accuracies reported for similar conversions. This approach may provide an important tool for formalism-independent evaluation. The unification of widely differing protein-protein interaction annotations of five corpora as part of Paper V made it possible to establish for the first time the magnitude of the effect of corpus on the measured performance of extraction methods, indicating serious problems in the comparability of current evaluations as well as providing one possible approach to addressing these issues.

The BioInfer corpus, introduced in Paper IV, has been critically important in supporting the research presented in this thesis. The corpus or previous, partial versions of its annotation have been applied not only in all of the studies described in Papers I, II, III and V, but also in numerous others. BioInfer was the first domain corpus to combine entity, relationship, and syntactic annotation in a single annotated resource and to break with the limited pairwise annotation strategy for protein-protein interactions (Paper IV), and it remains in many ways a unique resource for biomedical text mining. Its detailed annotation provides numerous unexplored opportunities for studying how protein-protein interactions are stated and how their extraction should best be approached.

One of the simultaneously frustrating and rewarding aspects of research is that it tends to open as many new questions as it answers. While the studies included in this thesis have contributed toward an understanding of domain parsing performance, challenges and adaptation techniques, the level of accuracy of the considered parsers on biomedical text remains con-

siderably lower than that expected on general English. Similarly, while the introduced conversion methodology shows promise for advancing formalism-independent parser evaluation and mapping between different representations, the generalizability of these results is yet to be established. Finally, protein-protein interaction extraction and, in particular, the more general task of extracting detailed, complex relationships between biomedical entities still hold many open questions, with the studies presented here providing necessary resources and indications of where problems lie rather than definite answers. I hope to have the opportunity to address some of these issues as future work.

References

Ahmed, S. T., Chidambaram, D., Davulcu, H., and Baral, C. (2005). IntEx: A syntactic role driven protein-protein interaction extractor for bio-medical text. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases (BioLINK'05)*, pages 54–61.

Airola, A., Pyysalo, S., Björne, J., , Pahikkala, T., Ginter, F., and Salakoski, T. (2008). A graph kernel for protein-protein interaction extraction. In *Proceedings of the ACL'08 Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP'08)*, pages 1–9. Association for Computational Linguistics.

Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., and Wang, X. (2008). Assisted curation: Does text mining really help? In *Proceedings of the Pacific Symposium on Biocomputing (PSB'08)*.

Alphonse, E., Aubin, S., Bessiéres, P., Bisson, G., Hamon, T., Laguarigue, S., Nazarenko, A., Manine, A.-P., Nédellec, C., Vetah, M. O. A., Poibeau, T., and Weissenbacher, D. (2004). Event-based information extraction for the biomedical domain: The Caderige project. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 43–49.

Ananiadou, S., Kell, D. B., and Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24:571–579.

Ananiadou, S. and McNaught, J., editors (2006). *Text Mining for Biology and Biomedicine*. Artech House Publishers.

Ananiadou, S. and Nenadic, G. (2006). Automatic terminology management in biomedicine. In Ananiadou, S. and McNaught, J., editors, *Text Mining for Biology and Biomedicine*, pages 67–97. Artech house.

Ando, R. K. (2007). BioCreative II gene mention tagging system at IBM Watson. In *Proceedings of the Second BioCreative Challenge Evaluation*, pages 101–103.

Andrade, M. A. and Valencia, A. (1998). Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–607.

Aubin, S. (2003). Evaluation comparative de deux analyseurs produisant des relations syntaxiques. In *Proceedings of the Workshop Traitement Automatique des Langues Naturelles (TALN)*, pages 67–76.

Aubin, S. (2005). LLL challenge - syntactic analysis guidelines. Technical report, LIPN, Université Paris Nord, Villetaneuse.

Aubin, S., Nazarenko, A., and Nédellec, C. (2005). Adapting a general parser to a sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 89–93.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T., and Hogue, C. W. (2001). BIND– the biomolecular interaction network database. *Nucleic Acids Reserarch*, 29(1):242–245.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. (2005). The universal protein resource (UniProt). *Nucleic Acids Research*, 33(Suppl. 1):D154–159.

Baumgartner, William A., J., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G., and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–48.

Björne, J., Pyysalo, S., Ginter, F., and Salakoski, T. (2008). How complex are complex protein-protein interactions? In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

Black, E., Abney, S., Flickenger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 306–311.

Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB'99)*, pages 60–67.

Blaschke, C. and Valencia, A. (2002). The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems*, 17(2):14–20.

Bloomfield, L. (1933). *Language*. Holt, Rinehart and Winston.

Bod, R. (2001). What is the minimal set of fragments that achieves maximal parse accuracy? In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, pages 66–73.

Bod, R. (2003). An efficient implementation of a new dop model. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 19–26.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Suppl. 1):D267–270.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370.

Bresnan, J. and Kaplan, R. (1982). Lexical-functional grammar: A formal system for grammatical representation. In *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP'02)*, pages 152–155.

Briscoe, T. and Carroll, J. (2002). Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1499–1504.

Bunescu, R. C. (2007). *Learning for Information Extraction: From Named Entity Recognition and Disambiguation To Relation Extraction*. PhD thesis, University of Texas at Austin.

Bunescu, R. C., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., and Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Bunescu, R. C. and Mooney, R. (2006). Subsequence kernels for relation extraction. In *Advances in Neural Information Processing Systems 18 (NIPS'06)*, pages 171–178.

Bunescu, R. C. and Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 724–731.

Carroll, J. E., Briscoe, E., and Sanfilippo, A. (1998). Parser evaluation: A survey and a new proposal. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*, pages 447–454.

Castaño, J. and Pustejovsky, J. (2005). Tagging with delayed disambiguation. In *Proceedings of the Fifth International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP'05)*, pages 285–287.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97)*.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, pages 132–139.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Chun, H.-W., Tsuruoka, Y., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., and Tsujii, J. (2006). Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'06)*, pages 4–15.

Clark, S. and Curran, J. (2007). Formalism-independent parser evaluation with ccg and depbank. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 248–255.

Clegg, A. B. (2008). *Computational-Linguistic Approaches to Biomedical Text Mining*. PhD thesis, University of London.

Clegg, A. B. and Shepherd, A. J. (2005). Evaluating and integrating treebank parsers on a biomedical corpus. In *Proceedings of the Association for Computational Linguistics Workshop on Software*.

Clegg, A. B. and Shepherd, A. J. (2007). Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics*, 8(1):24.

Cohen, A. M. and Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 6(1):57–71.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 184–191.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97)*, pages 16–23.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Collins, M. (2000). Discriminative reranking for natural language parsing. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML'00)*.

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

Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H., and Rebholz-Schuhmann, D. (2006). GOAnnotator: linking protein GO annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 1:19.

Covington, M. A. (1990). A dependency parser for variable-word-order languages. Technical Report AI-1990-01, University of Georgia.

Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems in Molecular Biology (ISMB'99)*, pages 77–86.

Culotta, A. and Sorensen, J. (2004). Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*, pages 423–429.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–611.

Deriviere, J., Hamon, T., and Nazarenko, A. (2006). A scalable and distributed NLP architecture for web document annotation. In *Proceedings of the Fifth International Conference on Natural Language Processing FinTAL'06*, pages 56–67.

Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. (2002). Mining MEDLINE: Abstracts, sentences, or phrases? In *Proceedings of the Pacific Symposium on Biocomputing (PSB'02)*, pages 326–337.

Ding, J., Berleant, D., Xu, J., and Fulmer, A. W. (2003). Extracting biochemical interactions from MEDLINE using a link grammar parser. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*, pages 467–471.

Dingare, S., Nissim, M., Finkel, J., Manning, C., and Grover, C. (2005). A system for identifying named entities in biomedical text: How results from two evaluations reflect on both the system and the evaluations. *Comparative and Functional Genomics*, 6(1-2):77–85.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) program: Tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 837–840.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., Pawson, T., and Hogue, C. W. (2003). PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4:11.

Entwisle, J. and Powers, D. (1998). The present use of statistics in the evaluation of nlp parsers. In *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Natural Language Learning (NeMLaP-CoNLL'98)*, pages 215–224.

Erkan, G., Özgür, A., and Radev, D. R. (2007). Semi-supervised classification for extracting protein interaction sentences using dependency parsing.

In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, pages 228–237.

Finkel, J., Dingare, S., Manning, C. D., Nissim, M., Alex, B., and Grover, C. (2005). Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics*, 6(Suppl. 1):S5.

Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C., and Sinclair, G. (2004). Exploiting context for biomedical entity recognition: From syntax to the web. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 88–91.

Franzén, K., Eriksson, G., Olsson, F., Asker, L., Lidén, P., and Cöster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics*, 4(67):49–61.

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

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl. 1):S74–S82.

Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'98)*, pages 707–718.

Fundel, K., Kuffner, R., and Zimmer, R. (2007). RelEx–Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Gaifman, H. (1965). Dependency systems and phrase-structure systems. *Information and Control*, 8:304–337.

Gaizauskas, R., Demetriou, G., Artymiuk, P. J., and Willett, P. (2003). Protein structures and information extraction from biological texts: The PASTA system. *Bioinformatics*, 19(1):135–143.

Galton, A. (2006). Processes as continuants (abstract). In *Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)*.

Ginter, F. (2007). *Towards Information Extraction in the Biomedical Domain: Methods and Resources*. PhD thesis, Turku Centre for Computer Science (TUCS).

Ginter, F., Boberg, J., Järvinen, J., and Salakoski, T. (2004a). New techniques for disambiguation in natural language and their application to biological text. *Journal of Machine Learning Research*, 5:605–621.

Ginter, F., Pyysalo, S., Björne, J., Heimonen, J., and Salakoski, T. (2007). BioInfer relationship annotation manual. Technical Report TR 806, Turku Centre for Computer Science (TUCS).

Ginter, F., Pyysalo, S., Boberg, J., Järvinen, J., and Salakoski, T. (2004b). Ontology-based feature transformations: A data-driven approach. In *Proceedings of the Fourth International Conference EsTAL 04, Alicante, Spain*, pages 279–290.

Ginter, F., Pyysalo, S., and Salakoski, T. (2005). Document classification using semantic networks with an adaptive similarity measure. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 204–211.

Giuliano, C., Lavelli, A., and Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 401–408.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, 274(5287):546–567.

Goodman, J. (1997). Probabilistic feature grammars. In *Proceedings of the Fourth International Workshop on Parsing Technologies (IWPT'97)*.

Grigoriev, A. (2003). On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research*, 31(14):4157–4161.

Grinberg, D., Lafferty, J., and Sleator, D. (1995). A robust parsing algorithm for link grammars. In *Proceedings of the Fourth International Workshop on Parsing Technologies (IWPT'95)*.

Grishman, R. (2001). Adaptive information extraction and sublanguage analysis. In *Proceedings of the IJCAI'01 Workshop on Adaptive Text Extraction and Mining*.

Grishman, R. (2003). Information extraction. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*, pages 545–559. Oxford University Press.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 466–471.

Grover, C., Carroll, J., and Briscoe, T. (1993). The Alvey natural language tools grammar (4th release). Technical Report 284, University of Cambridge.

Grover, C., Lapata, M., and Lascarides, A. (2005). A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering*, 11(1):27–65.

Grover, C. and Lascarides, A. (2001). Xml-based data preparation for robust deep parsing. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, pages 260–267.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hao, Y., Zhu, X., Huang, M., and Li, M. (2005). Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics*, 21(15):3294–3300.

Hara, T., Miyao, Y., and Tsujii, J. (2007). Evaluating impact of re-training a lexical disambiguation model on domain adaptation of an hpsg parser. In *Proceedings of the Tenth International Conference on Parsing Technologies (IWPT'07)*, pages 11–22.

Harris, Z. (1968). *Mathematical Structures of Language*. Wiley-Interscience.

Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 415–422.

Hatzivassiloglou, V., Duboué, P., A., and Rzhetsky, A. (2001). Disambiguating proteins, genes and RNA in text: A machine learning approach. *Bioinformatics*, 17(Suppl. 1):97–106.

Haverinen, K., Ginter, F., Pyysalo, S., and Salakoski, T. (2008). Accurate conversion of dependency parses: targeting the stanford scheme. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

Hays, D. G. (1964). Dependency theory: A formalism and some observations. *Language*, 40:511–525.

Heimonen, J., Pyysalo, S., Ginter, F., and Salakoski, T. (2008). Complex-to-pairwise mapping of biological relationships using a semantic network representation. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

Hersh, W., Buckley, C., Leone, T. J., and Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'94)*, pages 192–201.

Hersh, W. R., Bhupatiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., and Kraemer, D. F. (2004). TREC 2004 genomics track overview. In *Proceedings of the 13th Text Retrieval Conference (TREC'04)*.

Hirschman, L., Colosimo, M., Morgan, A., and Yeh, A. (2005a). Overview of BioCreAtIvE task 1B: Normalized gene lists. *BMC Bioinformatics*, 6(Suppl. 1):S11.

Hirschman, L., Morgan, A. A., and Yeh, A. S. (2002a). Rutabaga by any other name: extracting biological names. *Journal of Biomedical Informatics*, 35(4):247–259.

Hirschman, L., Park, J. C., Tsujii, J., Wong, L., and Wu, C. H. (2002b). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561.

Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005b). Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl. 1):S1.

Huang, L. (2008). Forest reranking: Discriminative parsing with non-local features. In *Proceedings of ACL-08: HLT*, pages 586–594.

Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K., and Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

Hunter, L. and Cohen, K. B. (2006). Biomedical language processing: What's beyond PubMed? *Molecular Cell*, 21(5):589–594.

Hunter, L., Lu, Z., Firby, J., Baumgartner, W. A., Johnson, H. L., Ogren, P. V., and Cohen, K. B. (2008). OpenDMAP: An open-source, ontology-driven concept analysis engine, with applications to capturing knowledge

regarding protein transport, protein interactions and cell-specific gene expression. *BMC Bioinformatics*, 9(78).

Jang, H., Lim, J., Lim, J.-H., Park, S.-J., Lee, K.-C., and Park, S.-H. (2006). Finding the evidence for protein-protein interactions from pubmed abstracts. *Bioinformatics*, 22(14):e220–226.

Jenssen, T.-K., Laegreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Suppl. 1):D428–432.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing*. Prentice-Hall.

Kakkonen, T. (2007). *Framework and Resources for Natural Language Parser Evaluation*. PhD thesis, University of Joensuu.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kaplan, R., Riezler, S., King, T. H., Maxwell III, J. T., Vasserman, A., and Crouch, R. (2004). Speed and accuracy in shallow and deep stochastic parsing. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HTL-NAACL'04)*, pages 97–104.

Karamanis, N., Seal, R., Lewin, I., McQuilton, P., Vlachos, A., Gasperin, C., Drysdale, R., and Briscoe, T. (2008). Natural Language Processing in aid of FlyBase curators. *BMC Bioinformatics*, 9:193.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th International Conference on Computational linguistics (COLING'90)*, pages 168–173.

Karopka, T., Scheel, T., Bansemer, S., and Glass, Ä. (2004). Automatic construction of gene relation networks using text mining and gene expression data. *Medical Informatics and the Internet in Medicine*, 29(2):169–183.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C., and Gama-Castro, S. (2002). The EcoCyc database. *Nucleic Acids Research*, 30(1):56–58.

Karttunen, L. (2007). Word play. *Computational Linguistics*, 33(4):443–467.

Katrenko, S. and Adriaans, P. (2006). Learning relations from biomedical corpora using dependency trees. In *Proceedings of the First Workshop on Knowledge Discovery and Emergent Complexity in BioInformatics (KDECB'06)*, pages 61–80.

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

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In Collier, N., Ruch, P., and Nazarenko, A., editors, *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 70–75.

Kim, S., Shin, S.-Y., Lee, I.-H., Kim, S.-J., Sriram, R., and Zhang, B.-T. (2008b). PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Research*, 36(Suppl. 2):W411–415.

Kim, S., Yoon, J., and Yang, J. (2008c). Kernel approaches for genic interaction extraction. *Bioinformatics*, 24(1):118–126.

Koike, A., Kobayashi, Y., and Takagi, T. (2003). Kinase pathway database: An integrated protein-kinase and nlp-based protein-interaction resource. *Genome Research*, 13:1241–1243.

Koike, A., Niwa, Y., and Takagi, T. (2005). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236.

Krallinger, M., Leitner, F., and Valencia, A. (2007). Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proceedings of the Second BioCreative Challenge Evaluation*, pages 41–54.

Krauthammer, M. and Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37:512–526.

Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A., and Ungar, L. (2004). Integrated annotation for biomedical information extraction. In *Proceedings of the HLT-NAACL 2004 Workshop on Linking Biological Literature, Ontologies and Databases (BioLINK'04)*, pages 61–68.

Laippala, V., Ginter, F., Pyysalo, S., and Salakoski, T. (2008). Resource-efficient construction of a full parser for Finnish nursing narratives. In

Proceedings of the First Louhi Conference on Text and Data Mining of Clinical Documents. To appear.

Lease, M. and Charniak, E. (2005). Parsing biomedical literature. In *Proceedings of the Second International Joint Conference on Natural Langage Processing (IJCNLP'05)*, pages 58–69.

Lehnert, W., Cardie, C., Fisher, D., McCarthy, J., Riloff, E., and Soderland, S. (1992). University of massachusetts: Muc-4 test results and analysis. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pages 151–158.

Leroy, G. and Chen, H. (2002). Filling preposition-based templates to capture information from medical abstracts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'02)*, pages 350–361.

Leroy, G., Chen, H., and Martinez, J. D. (2003). A shallow parser based on closed-class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, 36(3):145–158.

Leser, U. and Hakenberg, J. (2005). What makes a gene name? named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–369.

Levy, R. and Manning, C. (2004). Deep dependencies from context-free statistical parsers: Correcting the surface dependency approximation. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 327–334.

Lin, D. (1995). A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'95)*, pages 1420–1427.

Liu, H., Hu, Z.-Z., Zhang, J., and Wu, C. (2006). BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22(1):103–105.

Magerman, D. M. (1994). *Natural language parsing as statistical pattern recognition*. PhD thesis, Stanford University.

Magerman, D. M. (1995). Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 276–283.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 449–454.

Mathivanan, S., Periaswamy, B., Gandhi, T., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y., and Pandey, A. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, 7(Suppl. 5)(S19).

McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'06)*, pages 152–159.

McCray, A. T., Aronson, A. R., Browne, A. C., Rindflesch, T. C., Razi, A., and Srinivasan, S. (1993). UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association*, 81(2):184–194.

McDonald, D. M., Chen, H., Su, H., and Marshall, B. B. (2004). Extracting gene pathway relations using a hybrid grammar: The Arizona relation parser. *Bioinformatics*, 20(18):3370–3378.

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., and White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical IE. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 491–498.

McNaught, J. and Black, W. J. (2006). Information extraction. In Ananiadou, S. and McNaught, J., editors, *Text Mining for Biology and Biomedicine*, pages 143–177. Artech house.

Melli, G. (2007). Inductive approaches to the detection and classification of semantic relation mentions. Technical report, Simon Fraser School of Computing Science.

Melli, G., Ester, M., and Sarkar, A. (2007). Recognition of multi-sentence n-ary subcellular localization mentions in biomedical abstracts. In *Proceedings of the Second International Symposium on Languages in Biology and Medicine (LBM'07)*, pages 2.1–2.17.

Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schwartz, R., Stone, R., and Weischedel, R. (1998). Algorithms that learn to extract information - BBN: Description of the SIFT system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Mitkov, R., editor (2003). *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Mitsumori, T., Murata, M., Fukuda, Y., Doi, K., and Doi, H. (2006). Extracting protein-protein interaction information from biomedical text with SVM. *IEICE Transactions on Information and Systems*, E89-D(8):2464–2466.

Miyao, Y., Sætre, R., Sagae, K., Matsuzaki, T., and Tsujii, J. (2008). Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL'08)*, pages 46–54.

Miyao, Y., Sagae, K., and Tsujii, J. (2007). Towards framework-independent evaluation of deep linguistic parsers. In *Proceedings of the Grammar Engineering across Frameworks Workshop (GEAF'07)*.

Miyao, Y. and Tsujii, J. (2005). Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 83–90, Ann Arbor, Michigan. Association for Computational Linguistics.

Morgan, A. A. and Hirschman, L. (2007). Overview of BioCreative II gene normalization. In *Proceedings of the Second BioCreative Challenge Evaluation*, pages 101–103.

Mueller, E. T. (1987). *Daydreaming and Computation: A computer model of everyday creativity, learning, and emotions in the human stream of thought*. PhD thesis, University of California, Los Angeles.

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

Nédellec, C. (2005). Learning language in logic - genic interaction extraction challenge. In *Proceedings of the Learning Language in Logic Workshop (LLL'05)*.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL'07*, pages 915–932.

Nobata, C., Collier, N., and Tsujii, J. (2000). Comparison between tagged corpora for the named entity task. In *Proceedings of the ACL Workshop on Comparing Corpora*, pages 20–27.

Ohta, T., Miyao, Y., Ninomiya, T., Tsuruoka, Y., Yakushiji, A., Masuda, K., Takeuchi, J., Yoshida, K., Hara, T., Kim, J.-D., Tateisi, Y., and Tsujii, J. (2006). An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. In *Proceedings of COLING-ACL'06*, pages 17–20.

Ohta, T., Tateisi, Y., Mima, H., and Tsujii, J. (2002). GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference (HLT'02)*, pages 73–77.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161.

Pahikkala, T. (2008). *New Kernel Functions and Learning Methods for Text and Data Mining*. PhD thesis, Turku Centre for Computer Science (TUCS).

Pahikkala, T., Ginter, F., Boberg, J., Järvinen, J., and Salakoski, T. (2005a). Contextual weighting for support vector machines in literature mining: An application to gene versus protein name disambiguation. *BMC Bioinformatics*, 6(1):157.

Pahikkala, T., Pyysalo, S., Boberg, J., Järvinen, J., and Salakoski, T. (2008). Matrix representations, linear transformations, and kernels for natural language processing. *Machine Learning*. To appear.

Pahikkala, T., Pyysalo, S., Boberg, J., Mylläri, A., and Salakoski, T. (2005b). Improving the performance of bayesian and support vector classifiers in word sense disambiguation using positional information. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 90–97.

Pahikkala, T., Pyysalo, S., Ginter, F., Boberg, J., Järvinen, J., and Salakoski, T. (2005c). Kernels incorporating word positional information in natural language disambiguation tasks. In *Proceedings of the Eighteenth*

International Florida Artificial Intelligence Research Society Conference (FLAIRS'05), pages 442–447.

Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., and Salakoski, T. (2007). Learning to rank with pairwise regularized least-squares. In *SIGIR'07 Workshop on Learning to Rank for Information Retrieval*, pages 27–33.

Pahikkala, T., Tsivtsivadze, E., Boberg, J., and Salakoski, T. (2006). Graph kernels versus graph representations: A case study in parse ranking. In *Proceedings of the ECML-PKDD'06 workshop on Mining and Learning with Graphs (MLG'06)*.

Palakal, M., Stephens, M., Mukhopadhyay, S., Raje, R., and Rhodes, S. (2002). A multi-level text mining method to extract biological relationships. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB'02)*, pages 97–108.

Pallett, D. S., Garofolo, J. S., and Fiscus, J. G. (2000). Measurements in support of research accomplishments. *Communications of the ACM*, 43(2):75–79.

Park, J. C. (2001). Using combinatory categorial grammar to extract biomedical information. *IEEE Intelligent Systems*, 16(6):62–67.

Park, J. C., Kim, H. S., and Kim, J.-J. (2001). Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'01)*.

Park, J. C. and Kim, J.-J. (2006). Named entity recognition. In Ananiadou, S. and McNaught, J., editors, *Text Mining for Biology and Biomedicine*, pages 121–142. Artech house.

Peri, S. et al. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32(Suppl. 1):D497–501.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07)*, pages 404–411.

Phuong, T. M., Lee, D., and Lee, K. H. (2003). Learning rules to extract protein interactions from biomedical text. In *Proceedings of the seventh Pacific-Asia conference on knowledge discovery and data mining (PAKDD'03)*, pages 148–158.

Plake, C., Hakenberg, J., and Leser, U. (2005). Optimizing syntax patterns for discovering protein-protein interactions. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, pages 195–201.

Poggio, T. and Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society (AMS)*, 50(5):537–544.

Pollard, C. J. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(2):130–137.

Preiss, J. (2003). Using grammatical relations to compare parsers. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 291–298.

Proux, D., Rechenmann, F., and Juillard, L. (2000). A pragmatic information extraction strategy for gathering data on genetic interactions. In *Proceedings of Proceedings of the Eight International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, pages 279–285.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V., and Jacq, B. (1998). Detecting gene symbols and names in biological texts: A first step toward pertinent information extraction. *Genome Informatics*, 9:72–80.

Pustejovsky, J., Castaño, J., Zhang, J., Kotecki, M., and Cochran, B. (2002). Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'02)*, pages 362–373.

Pyysalo, S. (2003). Mining biomedical literature for protein-protein interactions using support vector machines. Master's thesis, University of Oulu.

Pyysalo, S., Ginter, F., Pahikkala, T., Boberg, J., Järvinen, J., Salakoski, T., and Koivula, J. (2004). Analysis of link grammar on biomedical dependency corpus targeted at protein-protein interactions. In *Proceedings of the International Workshop on Natural language Processing in Biomedicine and its Applications (JNLPBA)*, pages 15–21.

Pyysalo, S., Sætre, R., Tsujii, J., and Salakoski, T. (2008). Why biomedical relation extraction results are incomparable and what to do about it. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

Radford, A. (2004). *Minimalist Syntax: Exploring the Structure of English*. Cambridge University Press.

Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(R40).

Ratnaparkhi, A. (1997). A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP'97)*.

Rebholz-Schuhmann, D., Kirsch, H., and Couto, F. (2005). Facts from text – is text mining ready to deliver? *PLoS Biology*, 3(2):188–191.

Rebholz-Schuhmann, D., Marcel, S., Albert, S., Tolle, R., Casari, G., and Kirsch, H. (2004). Automatic extraction of mutations from Medline and cross-validation with OMIM. *Nucleic Acids Research*, 32(1):135–142.

Reynar, J. C. and Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pages 16–19.

Rifkin, R., Yeo, G., and Poggio, T. (2003). Regularized least-squares classification. In Suykens, J., Horvath, G., Basu, S., Micchelli, C., and Vandewalle, J., editors, *Advances in Learning Theory: Methods, Model and Applications*, chapter 7, pages 131–154.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann.

Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., and Romacker, M. (2006). An environment for relation mining over richly annotated corpora: The case of GENIA. In *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM'06)*, pages 68–75.

Rindflesch, T., Rajan, J., and Hunter, L. (2000a). Extracting molecular binding relationships from biomedical text. In *Proceedings of the Applied Natural Language Processing Conference of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL'00)*, pages 188–195.

Rindflesch, T., Tanabe, L., Weinstein, J. N., and Hunter, L. (2000b). EDGAR: Extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'00)*, pages 514–525.

Rindflesch, T. C., Hunter, L., and Aronson, A. R. (1999). Mining molecular binding terminology from biomedical text. In *Proceedings of the AMIA Annual Symposium*, pages 127–131.

Rosario, B. and Hearst, M. (2004). Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04)*, pages 430–437.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboué, P. A., Weng, W., Wilbur, W. J., Hatzivassiloglou, V., and Friedman, C. (2004). GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53.

Sætre, R., Sagae, K., and Tsujii, J. (2007). Syntactic features for protein-protein interaction extraction. In *Proceedings of the Second International Symposium on Languages in Biology and Medicine (LBM'07)*, pages 6.1–6.14.

Sagae, K., Miyao, Y., Matsuzaki, T., and Tsujii, J. (2008a). Challenges in mapping of syntactic representations for framework-independent parser evaluation. In *Proceedings of the ICGL'08 Workshop on Automated Syntatic Annotations for Interoperable Language Resources*.

Sagae, K., Miyao, Y., Saetre, R., and Tsujii, J. (2008b). Evaluating the effects of treebank size in a practical application for parsing. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 14–20.

Sagae, K. and Tsujii, J. (2007). Dependency parsing and domain adaptation with LR models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL'07*, pages 1044–1050.

Samuelsson, C. and Voutilainen, A. (1997). Comparing a linguistic and a stochastic tagger. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics (ACL'97)*, pages 246–253.

Sanchez-Graillet, O. and Poesio, M. (2007). Negation of protein protein interactions: analysis and extraction. *Bioinformatics*, 23(13):i424–432.

Schneider, G. (2007). *Hybrid Long-Distance Functional Dependency Parsing*. PhD thesis, University of Zurich.

Sekimizu, T., Park, H. S., and Tsujii, J. (1998). Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. *Genome Informatics*, 9:62–71.

Sekine, S. (1997). The domain dependence of parsing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pages 96–102.

Shen, L., Satta, G., and Joshi, A. (2007). Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 760–767.

Sleator, D. D. and Temperley, D. (1991). Parsing English with a Link Grammar. Technical Report CMU-CS-91-196, Carnegie Mellon University.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., and Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology*, 6(5):R46.

Smith, L., Rindflesch, T., and Wilbur, W. J. (2004). MedPost: A part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.

Spasic, I., Ananiadou, S., McNaught, J., and Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, 6(3):239–251.

Stapley, B. and Benoit, G. (2000). Biobibliometrics: Information retrieval and visualization from co-occurrences of gene names in medline abstracts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'00)*, pages 529–540.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122:957–968.

Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R., and Mostafa, J. (2001). Detecting gene relations from MEDLINE abstracts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'01)*, pages 483–495.

Suber, P. (2002). Open access to the scientific journal literature. *Journal of Biology*, 1(3).

Sun, C., Lin, L., Wang, X., and Guan, Y. (2007). Using maximum entropy model to extract protein-protein interaction information from biomedical literature. In *Proceedings of the Third International Conference on Intelligent Computing (ICIC'07)*, pages 730–737.

Sundheim, B. and Chinchor, N. (1995). Named entity task definition. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 319–332.

Sundheim, B. M. (1995). Overview of results of the muc-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 13–31.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 20:7–18.

Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557.

Szolovits, P. (2003). Adding a medical lexicon to an english parser. In *Proceedings of the 2003 AMIA Annual Symposium*, pages 639–643.

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

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97)*, pages 64–71.

Tateisi, Y., Yakushiji, A., Ohta, T., and Tsujii, J. (2005). Syntax annotation for the GENIA corpus. In *Proceedings of the Second International Joint Conference on Natural Langage Processing (IJCNLP'05)*, pages 222–227.

Temkin, J. M. and Gilder, M. R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19(16):2046–2053.

Tesnière, L. (1959). *Éléments de Syntaxe Structurale*. Klincksiek.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. (2000). Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'00)*, pages 538–549, Honolulu, HI.

Tomanek, K., Wermter, J., and Hahn, U. (2007). A reappraisal of sentence and token splitting for life science documents. In *Proceedings of the 12th International Medical Informatics Congress (MedInfo'07)*.

Torii, M., Kamboj, S., and Vijay-Shanker, K. (2003). An investigation of various information sources for classifying biological names. In *Proceedings of the ACL'03 Workshop on Natural Language Processing in the Biomedical Domain (BioNLP'03)*, pages 113–120.

Torii, M., Kamboj, S., and Vijay-Shanker, K. (2004). Using name-internal and contextual features to classify biological terms. *Journal of Biomedical Informatics*, 37:498–511.

Tsai, R. T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsian, J., Sung, T.-Y., and Hsu, W.-L. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7:92.

Tsivtsivadze, E., Pahikkala, T., Airola, A., Boberg, J., and Salakoski, T. (2008). A sparse regularized least-squares preference learning algorithm. In *Proceedings of the Tenth Scandinavian Conference on Artificial Intelligence (SCAI'08)*. To appear.

Tsivtsivadze, E., Pahikkala, T., Boberg, J., and Salakoski, T. (2007). Locality kernels for sequential data and their applications to parse ranking. *Applied Intelligence*. To appear.

Tsivtsivadze, E., Pahikkala, T., Pyysalo, S., Boberg, J., Mylläri, A., and Salakoski, T. (2005). Regularized least-squares for parse ranking. In *Proceedings of the Sixth International Symposium on Intelligent Data Analysis (IDA'05), Madrid, Spain*, pages 464–474.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Proceedings of the Panhellenic Conference on Informatics*, pages 382–392.

Turmo, J., Ageno, A., and Català, N. (2006). Adaptive information extraction. *ACM Computing Surveys*, 38(2):4.

Van Landeghem, S., Saeys, Y., De Baets, B., and Van de Peer, Y. (2008). Extracting protein-protein interactions from text using rich feature vectors and feature selection. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM'08)*. To appear.

Venter, C. J. et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Voutilainen, A. (1995). A syntax-based part-of-speech analyser. In *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, pages 157–164.

Voutilainen, A. (1997). Designing a (finite-state) parsing grammar. In Roche, E. and Schabes, Y., editors, *Finite-State Language Processing*, pages 283–303. MIT Press.

Voutilainen, A., Heikkilä, J., and Anttila, A. (1992). *Constraint grammar of English: A performance-oriented introduction.* University of Helsinki, Department of General Linguistics.

Wattarujeekrit, T., Shah, P., and Collier, N. (2004). PASBio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5(1):155.

Wilbur, J., Smith, L., and Tanabe, L. (2007). Biocreative 2 gene mention task. In *Proceedings of the Second BioCreative Challenge Evaluation*, pages 7–16.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80–83.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: The Database of Interacting Proteins. *Nucleic Acids Research*, 28(1):289–291.

Xiao, J., Su, J., Zhou, G., and Tan, C. (2005). Protein-protein interaction extraction: A supervised learning approach. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM'05)*, pages 51–59, Hinxton, UK.

Xuan, W., Watson, S. J., and Meng, F. (2007). Tagging sentence boundaries in biomedical literature. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'07)*, pages 186–195.

Yakushiji, A. (2006). *Relation Information Extraction Using Deep Syntactic Analysis.* PhD thesis, University of Tokyo.

Yakushiji, A., Miyao, Y., Ohta, T., Tateisi, Y., and Tsujii, J. (2006). Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, pages 284–292.

Yakushiji, A., Miyao, Y., Tateisi, Y., and Tsujii, J. (2005). Biomedical information extraction with predicate-argument structure patterns. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM'05)*, pages 60–69.

Yakushiji, A., Tateisi, Y., Miyao, Y., and Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. In *Proceedings of the Pacific Symposium on Biocomputing (PSB'01)*, pages 408–419.

Yang, Z., Lin, H., and Wu, B. (2007). BioPPIExtractor: A protein-protein interaction extraction system for biomedical literature. *Expert Systems with Applications*.

Yeh, A., Morgan, A., Colosimo, M., and Hirschman, L. (2005). BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl. 1):S2.

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

Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

Zhang, M., Zhang, J., and Su, J. (2006). Exploring syntactic features for relation extraction using a convolution tree kernel. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'06)*, pages 288–295.

Zhou, D., He, Y., and Kwoh, C. K. (2006). Extracting protein-protein interactions from the literature using the hidden vector state model. In *Proceedings of the Second International Workshop on Bioinformatics Research and Applications (IWBRA'06)*, pages 718–725.

Zhou, G., Shen, D., Zhang, J., Su, J., and Tan, S. (2005). Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*, 6(Suppl. 1):S7.

Zhou, G. and Su, J. (2004). Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 96–99.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. B. (2007). Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*.