What causes a causal relation? Detecting Causal Triggers in Biomedical Scientific Discourse

Claudiu Mihăilă and Sophia Ananiadou
The National Centre for Text Mining,
School of Computer Science,
The University of Manchester,
131 Princess Street, Manchester M1 7DN, UK
claudiu.mihaila@manchester.ac.uk
sophia.ananiadou@manchester.ac.uk

Abstract

Current domain-specific information extraction systems represent an important resource for biomedical researchers, who need to process vaster amounts of knowledge in short times. Automatic discourse causality recognition can further improve their workload by suggesting possible causal connections and aiding in the curation of pathway models. We here describe an approach to the automatic identification of discourse causality triggers in the biomedical domain using machine learning. We create several baselines and experiment with various parameter settings for three algorithms, i.e., Conditional Random Fields (CRF), Support Vector Machines (SVM) and Random Forests (RF). Also, we evaluate the impact of lexical, syntactic and semantic features on each of the algorithms and look at errors. The best performance of 79.35% F-score is achieved by CRFs when using all three feature types.

1 Introduction

The need to provide automated, efficient and accurate means of retrieving and extracting user-oriented biomedical knowledge has significantly increased according to the ever-increasing amount of knowledge pusblished daily in the form of research articles (Ananiadou and McNaught, 2006; Cohen and Hunter, 2008). Biomedical text mining has seen significant recent advancements in recent years (Zweigenbaum et al., 2007), including named entity recognition (Fukuda et al., 1998), coreference resolution (Batista-Navarro and Ananiadou, 2011; Savova et al., 2011) and relation (Miwa et al., 2009; Pyysalo et al., 2009) and event extraction (Miwa et al., 2012b; Miwa et al., 2012a). Using biomedical text mining technology, text can now be enriched via the addition of semantic metadata and thus can support tasks such as analysing molecular pathways (Rzhetsky et al., 2004) and semantic searching (Miyao et al., 2006).

However, more complex tasks, such as question answering and automatic summarisation, require the extraction of information that spans across several sentences, together with the recognition of relations that exist across sentence boundaries, in order to achieve high levels of performance.

The notion of *discourse* can be defined as a coherent sequence of clauses and sentences. These are connected in a logical manner by *discourse relations*, such as causal, temporal and conditional, which characterise how facts in text are related. In turn, these help readers infer deeper, more complex knowledge about the facts mentioned in the discourse. These relations can be either explicit or implicit, depending whether or not they are expressed in text using overt *discourse connectives* (also known as *triggers*). Take, for instance, the case in example (1), where the trigger *Therefore* signals a justification between the two sentences: because "a normal response to mild acid pH from PmrB requires both a periplasmic histidine and several glutamic acid residues", the authors believe that the "regulation of PmrB activity could involve protonation of some amino acids".

(1) In the case of PmrB, a normal response to mild acid pH requires not only a periplasmic histidine

but also several glutamic acid residues.
Therefore, regulation of PmrB activity may involve protonation of one or more of these amino acids.

Thus, by identifying this causal relation, search engines become able to discover relations between biomedical entities and events or between experimental evidence and associated conclusions. However, phrases acting as causal triggers in certain contexts may not denote causality in all cases. Therefore, a dictionary-based approach is likely to produce a very high number of false positives. In this paper, we explore several supervised machine-learning approaches to the automatic identification of triggers that actually denote causality.

2 Related Work

A large amount of work related to discourse parsing and discourse relation identification exists in the general domain, where researchers have not only identified discourse connectives, but also developed end-to-end discourse parsers (Pitler and Nenkova, 2009; Lin et al., 2012). Most work is based on the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), a corpus of lexically-grounded annotations of discourse relations.

Until now, comparatively little work has been carried out on causal discourse relations in the biomedical domain, although causal associations between biological entities, events and processes are central to most claims of interest (Kleinberg and Hripcsak, 2011). The equivalent of the PDTB for the biomedical domain is the BioDRB corpus (Prasad et al., 2011), containing 16 types of discourse relations, e.g., temporal, causal and conditional. The number of purely causal relations annotated in this corpus is 542. There are another 23 relations which are a mixture between causality and one of either background, temporal, conjunction or reinforcement relations. A slightly larger corpus is the BioCause (Mihăilă et al., 2013), containing over 850 manually annotated causal discourse relations in 19 full-text open-access journal articles from the infectious diseases domain.

Using the BioDRB corpus as data, some researchers explored the identification of discourse connectives (Ramesh et al., 2012). However, they do not distinguish between the types of discourse relations. They obtain the best F-score of 75.7% using CRF, with SVM reaching only 65.7%. These results were obtained by using only syntactic features, as sematic features were shown to lower the performance. Also, they prove that there exist differences in discourse triggers between the biomedical and general domains by training a model on the BioDRB and evaluating it against PDTB and vice-versa.

3 Methodology

In this section, we describe our data and the features of causal triggers. We also explain our evaluation methodology.

3.1 Data

The data for the experiments comes from the BioCause corpus. BioCause is a collection of 19 open-access full-text journal articles pertaining to the biomedical subdomain of infectious diseases, manually annotated with causal relationships. Two types of spans of text are marked in the text, namely causal triggers and causal arguments. Each causal relation is composed of three text-bound annotations: a trigger, a cause or evidence argument and an effect argument. Some causal relations have implicit triggers, so these are excluded from the current research.

Figure 1 shows an example of discourse causality from BioCause, marking the causal trigger and the two arguments with their respective relation. Named entities are also marked in this example.

BioCause contains 381 unique explicit triggers in the corpus, each being used, on average, only 2.10 times. The number decreases to 347 unique triggers when they are lemmatised, corresponding to an average usage of 2.30 times per trigger. Both count settings show the diversity of causality-triggering phrases that are used in the biomedical domain.

3.2 Features

Three types of features have been employed in the development of this causality trigger model, i.e., lexical, syntactic and semantic. These features are categorised and described below.

3.2.1 Lexical features

The lexical features are built from the actual tokens present in text. Tokenisation is performed by

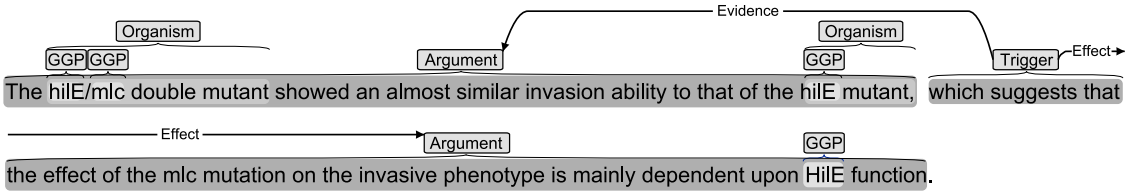

Figure 1: Causal relation in the BioCause.

the GENIA tagger (Tsuruoka et al., 2005) using the biomedical model. The first two features represent the token's surface expression and its base form.

Neighbouring tokens have also been considered. We included the token immediately to the left and the one immediately to the right of the current token. This decision is based on two observations. Firstly, in the case of tokens to the left, most triggers are found either at the beginning of the sentence (311 instances) or are preceded by a comma (238 instances). These two left contexts represent 69% of all triggers. Secondly, for the tokens to the right, almost 45% of triggers are followed by a determiner, such as *the*, *a* or *an*, (281 instances) or a comma (71 instances).

3.2.2 Syntactic features

The syntax, dependency and predicate argument structure are produced by the Enju parser (Miyao and Tsujii, 2008). Figure 2 depicts a partial lexical parse tree of a sentence which starts with a causal trigger, namely *Our results suggest that*. From the lexical parse trees, several types of features have been generated.

The first two features represent the part-of-speech and syntactic category of a token. For instance, the figure shows that the token *that* has the part-of-speech *IN*. These features are included due to the fact that either many triggers are lexicalised as an adverb or conjunction, or are part of a verb phrase. For the same reason, the syntactical category path from the root of the lexical parse tree to the token is also included. The path also encodes, for each parent constituent, the position of the token in its subtree, i.e., beginning (*B*), inside (*I*) or end (*E*); if the token is the only leaf node of the constituent, this is marked differently, using a *C*. Thus, the path of *that*, highlighted in the figure, is *I-S/I-VP/B-CP/C-CX*.

Secondly, for each token, we extracted the predicate argument structure and checked whether a relation exista between the token and the previous and following tokens. The values for this feature represent the argument number as allocated by Enju.

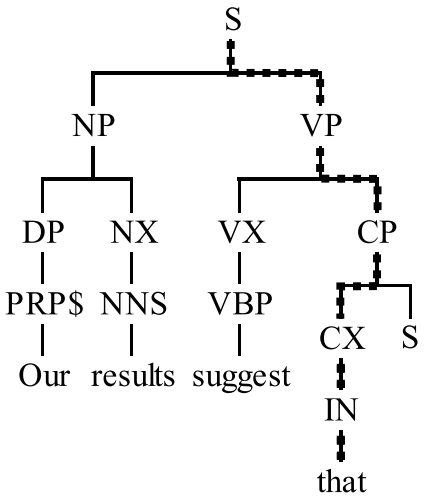

Figure 2: Partial lexical parse tree of a sentence starting with a causal trigger.

Thirdly, the ancestors of each token to the third degree are instantiated as three different features. In the case that such ancestors do not exist (i.e., the root of the lexical parse tree is less than three nodes away), a "none" value is given. For instance, the token *that* in Figure 2 has as its first three ancestors the constituents marked with *CX*, *CP* and *VP*.

Finally, the lowest common ancestor in the lexical parse tree between the current token and its left neighbour has been included. In the example, the lowest common ancestor for *that* and *suggest* is *VP*.

These last two feature types have been produced on the observation that the lowest common ancestor for all tokens in a causal trigger is S or VP in over 70% of instances. Furthermore, the percentage of cases of triggers with V or ADV as lowest common ancestor is almost 9% in each case. Also, the aver-

age distance to the lowest common ancestor is 3.

3.2.3 Semantic features

We have exploited several semantic knowledge sources to identify causal triggers more accurately, as a mapping to concepts and named entities acts as a back-off smoothing, thus increasing performance.

One semantic knowledge source is the BioCause corpus itself. All documents annotated for causality in BioCause had been previously manually annotated with biomedical named entity and event information. This was performed in the context of various shared tasks, such as the BioNLP 2011 Shared Task on Infectious Diseases (Pyysalo et al., 2011). We therefore leverage this existing information to add another semantic layer to the model. Moreover, another advantage of having a gold standard annotation is the fact that it is now possible to separate the task of automatic causal trigger recognition from automatic named entity recognition and event extraction. The named entity and event annotation in the BioCause corpus is used to extract information about whether a token is part of a named entity or event trigger. Furthermore, the type of the named entity or event is included as a separate feature.

The second semantic knowledge source is WordNet (Fellbaum, 1998). Using this resource, the hypernym of every token in the text has been included as a feature. Only the first sense of every token has been considered, as no sense disambiguation technique has been employed.

Finally, tokens have been linked to the Unified Medical Language System (UMLS) (Bodenreider, 2004) semantic types. Thus, we included a feature to say whether a token is part of a UMLS type and another for its semantic type if the previous is true.

3.3 Experimental setup

We explored with various machine learning algorithms and various settings for the task of identifying causal triggers.

On the one hand, we experimented with CRF (Lafferty et al., 2001), a probabilistic modelling framework commonly used for sequence labelling tasks. In this work, we employed the CRFSuite implementation[1].

[1] http://www.chokkan.org/software/crfsuite

On the other hand, we modelled trigger detection as a classification task, using Support Vector Machines and Random Forests. More specifically, we employed the implementation in Weka (Hall et al., 2009; Witten and Frank, 2005) for RFs, and LibSVM (Chang and Lin, 2011) for SVMs.

4 Results and discussion

Several models have been developed and 10-fold cross-evaluated to examine the complexity of the task, the impact of various feature types (lexical, syntactic, semantic). Table 1 shows the performance evaluation of baseline systems and other classifiers. These are described in the following subsections. It should be noted that the dataset is highly skewed, with a ratio of positive examples to negative examples of approximately 1:52.

	Classifier	P	R	F_1
Baseline	*Dict*	8.36%	100%	15.43%
	Depend	7.51%	76.66%	13.68%
	Dict+Depend	14.30%	75.33%	24.03%
2-way	CRF	**89.29%**	**73.53%**	**79.35%**
	SVM	81.62%	61.05%	69.85%
	RandFor	78.16%	66.96%	72.13%
3-way	CRF	**89.13%**	**64.04%**	**72.87%**
	SVM	74.21%	56.82%	64.36%
	RandFor	73.80%	60.95%	66.76%

Table 1: Performance of various classifiers in identifying causal connectives

4.1 Baseline

Several baselines have been devised. The first baseline is a dictionary-based heuristic, named *Dict*. A lexicon is populated with all annotated causal triggers and then this is used to tag all instances of its entries in the text as connectives. The precision of this heuristic is very low, 8.36%, which leads to an F-score of 15.43%, considering the recall is 100%. This is mainly due to triggers which are rarely used as causal triggers, such as *and*, *by* and *that*.

Building on the previously mentioned observation about the lowest common ancestor for all tokens in a causal trigger, we built a baseline system that checks all constituent nodes in the lexical parse tree for the S, V, VP and ADV tags and marks them as

causal triggers. The name of this system is *Depend*. Not only does *Depend* obtain a lower precision than *Dict*, but it also performs worse in terms of recall. The F-score is 13.68%, largely due to the high number of intermediate nodes in the lexical parse tree that have VP as their category.

The third baseline is a combination of *Dict* and *Depend*: we consider only constituents that have the necessary category (S, V, VP or ADV) and include a trigger from the dictionary. Although the recall decreases slightly, the precision increases to almost twice that of both *Dict* and *Depend*. This produces a much better F-score of 24.03%.

4.2 Sequence labelling task

As a sequence labelling task, we have modelled causal trigger detection as two separate tasks. Firstly, each trigger is represented in the B-I-O format (further mentioned as the 3-way model). Thus, the first word of every trigger is tagged as B (*begin*), whilst the following words in the trigger are tagged as I (*inside*). Non-trigger words are tagged as O (*outside*).

The second model is a simpler version of the previous one: it does not distinguish between the first and the following words in the trigger. In other words, each word is tagged either as being part of or outside the trigger, further known as the 2-way model. Hence, a sequence of contiguous tokens marked as part of a trigger form one trigger.

CRF performs reasonably well in detecting causal triggers. In the 3-way model, it obtains an F-score of almost 73%, much better than the other algorithms. It also obtains the highest precision (89%) and recall (64%). However, in the 2-way model, CRF's performance is slightly lower than that of Random Forests, achieving only 79.35%. Its precision, on the other hand, is the highest in this model. The results from both models were obtained by combining features from all three feature categories.

Table 2 show the effect of feature types on both models of CRFs. As can be observed, the best performances, in terms of F-score, including the previously mentioned ones, are obtained when combining all three types of features, i.e., lexical, syntactic and semantic. The best precision and recall, however, are not necessarily achieved by using all three feature types. In the two-way model, the best precision is obtained by using the syntactic features only, reaching over 92%, almost 3% higher than when all three feature types are used. In the three-way model, syntactic and semantic features produce the best recall (almost 65%), which is just under 1% higher than the recall when all features are used.

	Features	**P**	**R**	**F_1**
2-way	Lex	88.99%	67.09%	73.59%
	Syn	**92.20%**	68.68%	75.72%
	Sem	87.20%	63.30%	69.36%
	Lex+Syn	87.76%	73.29%	78.73%
	Lex+Sem	89.54%	69.10%	75.61%
	Syn+Sem	87.48%	72.62%	78.13%
	Lex+Syn+Sem	89.29%	**73.53%**	**79.35%**
3-way	Lex	85.87%	56.34%	65.18%
	Syn	87.62%	61.44%	70.22%
	Sem	80.78%	51.43%	59.39%
	Lex+Syn	87.80%	63.04%	72.59%
	Lex+Sem	85.50%	58.11%	66.80%
	Syn+Sem	84.83%	**64.94%**	72.41%
	Lex+Syn+Sem	**89.13%**	64.04%	**72.87%**

Table 2: Effect of feature types on the sequence labelling task.

4.3 Classification task

As a classification task, an algorithm has to decide whether a token is part of a trigger or not, similarly to the previous two-way subtask in the case of CRF.

Firstly, we have used RF for the classification task. Various parameter settings regarding the number of constructed trees and the number of random features have been explored.

The effect of feature types on the performance of RF is shown in Table 3. As can be observed, the best performance is obtained when combining lexical and semantic features. Due to the fact that causal triggers do not have a semantic mapping to concepts in the named entity and UMLS annotations, the trees in the random forest classifier can easily produce rules that distinguish triggers from non-triggers. As such, the use of semantic features alone produce a very good precision of 84.34%. Also, in all cases where semantic features are combined with other feature types, the precision increases by 0.5% in the case of lexical features and 3.5% in the case of syntactic features. However, the recall of semantic fea-

tures alone is the lowest. The best recall is obtained when using only lexical features.

Features	P	R	F_1
Lex	78.47%	**68.30%**	73.03%
Syn	68.19%	62.36%	65.15%
Sem	**84.34%**	56.83%	67.91%
Lex+Syn	77.11%	65.92%	71.09%
Lex+Sem	79.10%	67.91%	**73.08%**
Syn+Sem	71.83%	64.45%	67.94%
Lex+Syn+Sem	77.98%	67.31%	72.25%

Table 3: Effect of feature types on Random Forests.

Secondly, we explored the performance of SVMs in detecting causal triggers. We have experimented with two kernels, namely polynomial (second degree) and radial basis function (RBF) kernels. For each of these two kernels, we have evaluated various combinations of parameter values for cost and weight. Both these kernels achieved similar results, indicating that the feature space is not linearly separable and that the problem is highly complex.

The effect of feature types on the performance of SVMs is shown in Table 4. As can be observed, the best performance is obtained when combining the lexical and semantic feature types (69.85% F-score). The combination of all features produces the best precision, whilst the best recall is obtained by combining lexical and semantic features.

Features	P	R	F_1
Lex	80.80%	60.94%	69.47%
Syn	82.94%	55.60%	66.57%
Sem	85.07%	56.51%	67.91%
Lex+Syn	86.49%	53.63%	66.81%
Lex+Sem	81.62%	**61.05%**	**69.85%**
Syn+Sem	84.49%	55.31%	66.85%
Lex+Syn+Sem	**87.70%**	53.96%	66.81%

Table 4: Effect of feature types on SVM.

4.4 Error analysis

As we expected, the majority of errors arise from sequences of tokens which are only used infrequently as non-causal triggers. This applies to 107 trigger types, whose number of false positives (FP) is higher than the number of true positives (TP). In fact, 64 trigger types occur only once as a causal instance, whilst the average number of FPs for these types is 14.25. One such example is *and*, for which the number of non-causal instances (2305) is much greater than that of causal instances (1). Other examples of trigger types more commonly used as causal triggers, are *suggesting* (9 TP, 54 FP), *indicating* (8 TP, 41 FP) and *resulting in* (6 TP, 14 FP). For instance, example (2) contains two mentions of *indicating*, but neither of them implies causality.

(2) Buffer treated control cells showed intense green staining with syto9 (*indicating* viability) and a lack of PI staining (*indicating* no dead/dying cells or DNA release).

5 Conclusions and Future Work

We have presented an approach to the automatic identification of triggers of causal discourse relations in biomedical scientific text. The task has proven to be a highly complex one, posing many challenges. Shallow approaches, such as dictionary matching and lexical parse tree matching, perform very poorly, due to the high ambiguity of causal triggers (with F-scores of approximately 15% each and 24% when combined). We have explored various machine learning algorithms that automatically classify tokens into triggers or non-triggers and we have evaluated the impact of multiple lexical, syntactic and semantic features. The performance of SVMs prove that the task of identifying causal triggers is indeed complex. The best performing classifier is CRF-based and combines lexical, syntactical and semantical features in order to obtain an F-score of 79.35%.

As future work, integrating the causal relations in the BioDRB corpus is necessary to check whether a data insufficiency problem exists and, if so, estimate the optimal amount of necessary data. Furthermore, evaluations against the general domain need to be performed, in order to establish any differences in expressing causality in the biomedical domain. One possible source for this is the PDTB corpus. A more difficult task that needs attention is that of identifying implicit triggers. Finally, our system needs to be extended in order to identify the two arguments of

causal relations, the cause and effect, thus allowing the creation of a complete discourse causality parser.

Acknowledgements

This work was partially funded by the Engineering and Physical Sciences Research Council [grant number EP/P505631/1].

References

Sophia Ananiadou and John McNaught, editors. 2006. *Text Mining for Biology And Biomedicine*. Artech House, Inc.

Riza Theresa B. Batista-Navarro and Sophia Ananiadou. 2011. Building a coreference-annotated corpus from the domain of biochemistry. In *Proceedings of BioNLP 2011*, pages 83–91.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Kevin Bretonnel Cohen and Lawrence Hunter. 2008. Getting started in text mining. *PLoS Computational Biology*, 4(1):e20, 01.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ken-ichiro Fukuda, Tatsuhiko Tsunoda, Ayuchi Tamura, and Toshihisa Takagi. 1998. Toward information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 707, pages 707–718.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November.

Samantha Kleinberg and George Hripcsak. 2011. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, 44(6):1102 – 1112.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2012. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, FirstView:1–34, 10.

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

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2009. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78(12):e39–e46, June.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012a. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

Makoto Miwa, Paul Thompson, John McNaught, Douglas B. Kell, and Sophia Ananiadou. 2012b. Extracting semantically enriched events from biomedical literature. *BMC Bioinformatics*, 13:108.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic HPSG parsing. *Computational Linguistics*, 34(1):3580, March.

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

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *ACL/AFNLP (Short Papers)*, pages 13–16.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *In Proceedings of the 6th International Conference on language Resources and Evaluation (LREC)*, pages 2961–2968.

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

Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2011. Overview of the infectious diseases (ID) task of BioNLP shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 26–35, Portland, Oregon, USA, June. Association for Computational Linguistics.

Polepalli Balaji Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington, and Hong Yu. 2012. Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association*.

Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Ariel Pablo Duboué, Wubin Weng, W.John Wilbur, Vasileios Hatzivassiloglou, and Carol Friedman. 2004. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43 – 53.

Guergana K Savova, Wendy W Chapman, Jiaping Zheng, and Rebecca S Crowley. 2011. Anaphoric relations in the clinical narrative: corpus creation. *Journal of the American Medical Informatics Association*, 18(4):459–465.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, volume 3746 of *LNCS*, pages 382–392. Springer-Verlag, Volos, Greece, November.

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

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. 2007. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375.