Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences

Oana Frunza and Diana Inkpen
School of Information Technology and Engineering
University of Ottawa Ottawa, ON, Canada, K1N 6N5
{ofrunza,diana}@site.uottawa.ca

Abstract

This paper describes our study on identifying semantic relations that exist between diseases and treatments in biomedical sentences. We focus on three semantic relations: *Cure*, *Prevent*, and *Side Effect*. The contributions of this paper consists in the fact that better results are obtained compared to previous studies and the fact that our research settings allow the integration of biomedical and medical knowledge. We obtain 98.55% F-measure for the *Cure* relation, 100% F-measure for the *Prevent* relation, and 88.89% F-measure for the *Side Effect* relation.

1 Introduction

Research in the fields of life-science and biomedical domain has been the focus of the Natural Language Processing (NLP) and Machine Learning (ML) community for some time now. This trend goes very much inline with the direction the medical healthcare system is moving to: the electronic world. The research focus of scientists that work in the filed of computational linguistics and life science domains also followed the trends of the medicine that is practiced today, an Evidence Based Medicine (EBM). This new way of medical practice is not only based on the experience a healthcare provider acquires as time passes by, but on the latest discoveries as well. We live in an information explosion era where it is almost impossible to find that piece of relevant information that we need. With easy and cheep access to disk-space we sometimes even find challenging to find our stored local documents. It should come to no surprise that the global trend in domains like biomedicine and not only is to rely on technology to identify and upraise information. The amount of publications and research that is indexed in the life-science domain grows almost exponentially (Hunter and Cohen (2006) making the task of finding relevant information, a hard and challenging task for NLP research.

The search for information in the life-science domain is not only the focus of researchers that work in these fields, but the focus of laypeople as well. Studies reveal that people are searching the web for medical-related articles to be better informed about their health. Ginsberg *et al.* (2009) show how a new outbreak of the influenza virus can be detected from search engine query data.

The aim of this paper is to show which NLP and ML techniques are suitable for the task of identifying semantic relations between diseases and treatments in short biomedical texts. The value of our work stands in the results we obtain and the new feature representation techniques.

2 Related Work

The most relevant work for our study is the work of Rosario and Hearst (2004). The authors of this paper are the ones that created and distributed the data set used in our research. The data set is annotated with disease and treatments entities and with 8 semantic relations between diseases and treatments. The main focus of their work is on entity recognition – the task of identifying entities, diseases and treatments in biomedical text sentences. The authors use Hidden Markov Models and maximum entropy models to perform both the task of entity recognition and of relation discrimination. Their representation techniques are based on words in context, part-of-speech information, phrases, and terms from MeSH[1], a medical lexical knowledge-base. Compared to previous work, our research is focused

[1] http://www.nlm.nih.gov/mesh/meshhome.html

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 91–98,
Uppsala, Sweden, 15 July 2010. ©2010 Association for Computational Linguistics

on different representation techniques, different classification models, and most importantly in obtaining improved results without using the annotations of the entities (new data will not have them). In previous research, the best results were obtained when the entities involved in the relations were identified and used as features.

The biomedical literature contains a wealth of work on semantic relation extraction, mostly focused on more biology-specific tasks: *subcellular-location* (Craven 1999), *gene-disorder* association (Ray and Craven 2001), and *diseases and drugs* relations (Srinivasan and Rindflesch 2002, Ahlers et *al.*, 2007).

Text classification techniques combined with a Naïve Bayes classifier and relational learning algorithms are methods used by Craven (1999). Hidden Markov Models are used in Craven (2001), but similarly to Rosario and Hearst (2004), the research focus was entity recognition.

A context based approach using MeSH term co-occurrences are used by Srinivasan and Rindflesch (2002) for relationship discrimination between diseases and drugs.

A lot of work is focused on building rules used to extract relation. Feldman et *al.* (2002) use a rule-based system to extract relations that are focused on genes, proteins, drugs, and diseases. Friedman et *al.* (2001) go deeper into building a rule-based system by hand-crafting a semantic grammar and a set of semantic constraints in order to recognize a range of biological and molecular relations.

3 Task and Data Sets

Our task is focused on identifying disease-treatment relations in sentences. Three relations: *Cure*, *Prevent*, and *Side Effect*, are the main objective of our work. We are tackling this task by using techniques based on NLP and supervised ML techniques. We decided to focus on these three relations because these are the ones that are better represented in the original data set and in the end will allow us to draw more reliable conclusions. Also, looking at the meaning of all relations in the original data set, the three that we focus on are the ones that could be useful for wider research goals and are the ones that really entail relations between two entities. In the supervised ML settings the amount of training data is a factor that influences the performance; support for this stands not only in the related work performed on the same data set, but in the research literature as well. The aim of this paper is to focus on few relations of interest and try to identify what predictive model and what representation techniques bring the best results of identifying semantic relations in short biomedical texts. We mostly focused on the value that the research can bring, rather than on an incremental research.

As mentioned in the previous section, the data set that we use to run our experiments is the one of Rosario and Hearst (2004). The entire data set is collected from Medline[2] 2001 abstracts. Sentences from titles and abstracts are annotated with entities and with 8 relations, based only on the information present in a certain sentence. The first 100 titles and 40 abstracts from each of the 59 Medline 2001 files were used for annotation. Table 1, presents the original data set, as published in previous research. The numbers in parenthesis represent the training and test set sizes.

Relationship	Definition and Example
Cure 810 (648, 162)	TREAT cures DIS *Intravenous immune globulin for recurrent spontaneous abortion*
Only DIS 616 (492, 124)	TREAT not mentioned *Social ties and susceptibility to the common cold*
Only TREAT 166 (132, 34)	DIS not mentioned *Flucticasome propionate is safe in recommended doses*
Prevent 63 (50, 13)	TREAT prevents the DIS *Statins for prevention of stroke*
Vague 36 (28, 8)	Very unclear relationship *Phenylbutazone and leukemia*
Side Effect 29 (24, 5)	DIS is a result of a TREAT *Malignant mesodermal mixed tumor of the uterus following irradiation*
NO Cure 4 (3, 1)	TREAT does not cure DIS *Evidence for double resistance to permethrin and malathion in head lice*
Total relevant: 1724 (1377, 347)	
Irrelevant 1771 (1416, 355)	Treat and DIS not present *Patients were followed up for 6 months*
Total: 3495 (2793, 702)	

Table 1. Original data set.

From this original data set, the sentences that are annotated with *Cure*, *Prevent*, *Side Effect*, *Only DIS*, *Only TREAT*, and *Vague* are the ones that used in our current work. While our main focus is on the *Cure*, *Prevent*, and *Side Effect*, we also run experiments for all relations such that a direct comparison with the previous work is done.

[2] http://medline.cos.com/

Table 2 describes the data sets that we created from the original data and used in our experiments. For each of the relations of interest we have 3 labels attached: *Positive*, *Negative*, and *Neutral*. The *Positive* label is given to sentences that are annotated with the relation in question in the original data; the *Negative* label is given to the sentences labeled with *Only DIS* and *Only TREAT* classes in the original data; *Neutral* label is given to the sentences annotated with *Vague* class in the original data set.

Relation	**Train**		
	Positive	**Negative**	**Neutral**
Cure	554	531	25
Prevent	42	531	25
SideEffect	20	531	25
Relation		**Test**	
	Positive	**Negative**	**Neutral**
Cure	276	266	12
Prevent	21	266	12
SideEffect	10	266	12

Table 2. Our data sets[3].

4 Methodology

The experimental settings that we follow are adapted to the domain of study (we integrate additional medical knowledge), yielding for the methods to bring improved performance.

The challenges that can be encountered while working with NLP and ML techniques are: finding the suitable model for prediction – since the ML field offers a suite of predictive models (algorithms), the task of finding the suitable one relies heavily on empirical studies and knowledge expertise; and finding the best data representation – identifying the right and sufficient features to represent the data is a crucial aspect. These challenges are addressed by trying various predictive algorithms based on different learning techniques, and by using various textual representation techniques that we consider suitable.

The task of identifying the three semantic relations is addressed in three ways:

Setting 1: build three models, each focused on one relation that can distinguish sentences that contain the relation – *Positive* label, from other sentences that are neutral – *Neutral* label, and from sentences that do not contain relevant information – *Negative* label;

Setting 2: build three models, each focused on one relation that can distinguish sentences that contain the relation from sentences that do not contain any relevant information. This setting is similar to a two-class classification task in which instances are labeled either with the relation in question – *Positive* label, or with non-relevant information – *Negative* label;

Setting 3: build one model that distinguishes the three relations – a three-way classification task where each sentence is labeled with one of the semantic relations, using the data with all the *Positive* labels.

The first set of experiments is influenced by previous research done by Koppel and Schler (2005). The authors claim that for polarity learning "neutral" examples help the learning algorithms to better identify the two polarities. Their research was done on a corpus of posts to chat groups devoted to popular U.S. television and posts to shopping.com's product evaluation page.

As classification algorithms, a set of 6 representative models: decision-based models (Decision trees – J48), probabilistic models (Naïve Bayes and complement Naïve Bayes (CNB), which is adapted for imbalanced class distribution), adaptive learning (AdaBoost), linear classifier (support vector machine (SVM) with polynomial kernel), and a classifier, ZeroR, that always predicts the majority class in the training data used as a baseline. All classifiers are part of a tool called Weka[4].

As representation technique, we rely on features such as the words in the context, the noun and verb-phrases, and the detected biomedical and medical entities. In the following subsections, we describe all the representation techniques that we use.

4.1 Bag-of-words representation

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which the features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped into this feature representation by giving values to each feature for a certain instance. Two feature value representations are the most commonly used for the BOW representation: binary feature values – the value

[3] The number of sentences available for download is not the same as the ones from the original data set, published in Rosario and Hearst ('04).

[4] http://www.cs.waikato.ac.nz/ml/weka/

of a feature is 1 if the feature is present in the instance and 0 otherwise, or frequency feature values – the feature value is the number of times it appears in an instance, or 0 if it did not appear.

Taking into consideration the fact that an instance is a sentence, the textual information is relatively small. Therefore a frequency value representation is chosen. The difference between a binary value representation and a frequency value representation is not always significant, because sentences tend to be short. Nonetheless, if a feature appears more than once in a sentence, this means that it is important and the frequency value representation captures this aspect.

The selected features are words (not lemmatized) delimited by spaces and simple punctuation marks: *space, (,) , [,] , . , ' , _* that appeared at least three times in the training collection and contain at least an alpha-numeric character, are not part of an English list of stop words[5] and are longer than three characters. Stop words are function words that appear in every document (e.g., *the, it, of, an*) and therefore do not help in classification. The frequency threshold of three is commonly used for text collections because it removes non-informative features and also strings of characters that might be the result of a wrong tokenization when splitting the text into words. Words that have length of one or two characters are not considered as features because of two reasons: possible incorrect tokenization and problems with very short acronyms in the medical domain that could be highly ambiguous (could be a medical acronym or an abbreviation of a common word).

4.2 NLP and biomedical concepts representation

The second type of representation is based on NLP information – noun-phrases, verb-phrases and biomedical concepts (Biomed). In order to extract this type of information from the data, we used the Genia[6] tagger. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as Medline abstracts.

Figure 1 presents an output example by the Genia tagger for the sentence: "*Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin.*". The tag O stands for Outside, B for Beginning, and I for Inside.

Figure 1. Example of Genia tagger output

```
Inhibition   Inhibition       NN    B-NP   O
of           of               IN    B-PP   O
NF-kappaB NF-kappaB           NN    B-NP   B-protein
activation   activation       NN    I-NP   O
reversed     reverse          VBD   B-VP   O
the          the              DT    B-NP   O
anti-apoptotic anti-apoptotic JJ    I-NP   O
effect       effect           NN    I-NP   O
of           of               IN    B-PP   O
isochamaejasmin isochamaejasmin NN B-NP    O
.            .                .     O      O
```

The noun-phrases and verb-phrases identified by the tagger are considered as features for our second representation technique. The following preprocessing steps are applied before defining the set of final features: remove features that contain only punctuation, remove stop-words, and consider valid features only the lemma-based forms of the identified noun-phrases, verb-phrases and biomedical concepts. The reason to do this is because there are a lot of inflected forms (*e.g.,* plural forms) for the same word and the lemmatized form (the base form of a word) will give us the same base form for all the inflected forms.

4.3 Medical concepts (UMLS) representation

In order to work with a representation that provides features that are more general than the words in the abstracts (used in the BOW representation), we also used the unified medical language system[7] (here on UMLS) concept representations. UMLS is a knowledge source developed at the U.S. National Library of Medicine (here on NLM) and it contains a meta-thesaurus, a semantic network, and the specialist lexicon for biomedical domain. The meta-thesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts. UMLS contains over 1 million medical concepts, and over 5 million concept names which are hierarchical organized. Each unique concept that is present in the thesaurus has associated multiple text strings variants (slight morphological variations of the concept). All concepts are assigned at least one semantic type from the semantic network providing a generalization of the existing relations between concepts. There are 135 semantic types in the knowledge base linked through 54 relationships.

[5] http://www.site.uottawa.ca/~diana/csi5180/StopWords

[6] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

[7] http://www.nlm.nih.gov/pubs/factsheets/umls.html

In addition to the UMLS knowledge base, NLM created a set of tools that allow easier access to the useful information. MetaMap[8] is a tool created by NLM that maps free text to medical concepts in the UMLS, or equivalently, it discovers meta-thesaurus concepts in text. With this software, text is processed through a series of modules that in the end will give a ranked list of all possible concept candidates for a particular noun-phrase. For each of the noun phrases that the system finds in the text, variant noun phrases are generated. For each of the variant noun phrases, candidate concepts (concepts that contain the noun phrase variant) from the UMLS meta-thesaurus are retrieved and evaluated. The retrieved concepts are compared to the actual phrase using a fit function that measures the text overlap between the actual phrase and the candidate concept (it returns a numerical value). The best of the candidates are then organized according to the decreasing value of the fit function. We used the top concept candidate for each identified phrase in an abstract as a feature. Figure 2 presents an example of the output of the MetaMap system for the phrase "*to an increased risk*". The information presented in the brackets, the semantic type, "Qualitative Concept, Quantitative Concept" for the candidate with the fit function value 861 is the feature used for our UMLS representation.

Figure 2. Example of MetaMap system output

Meta Candidates (6)
861 Risk [Qualitative Concept, Quantitative Concept]
694 Increased (Increased (qualifier value)) [Functional Concept]
623 Increase (Increase (qualifier value)) [Functional Concept]
601 Acquired (Acquired (qualifier value)) [Temporal Concept]
601 Obtained (Obtained (attribute)) [Functional Concept]
588 Increasing (Increasing (qualifier value)) [Functional Concept]

Another reason to use a UMLS concept representation is the ***concept drift*** phenomenon that can appear in a BOW representation. Especially in the medical domain texts, this is a frequent problem as stated by Cohen *et al.* (2004). New articles that publish new research on a certain topic bring with them new terms that might not match the ones that were seen in the training process in a certain moment of time.

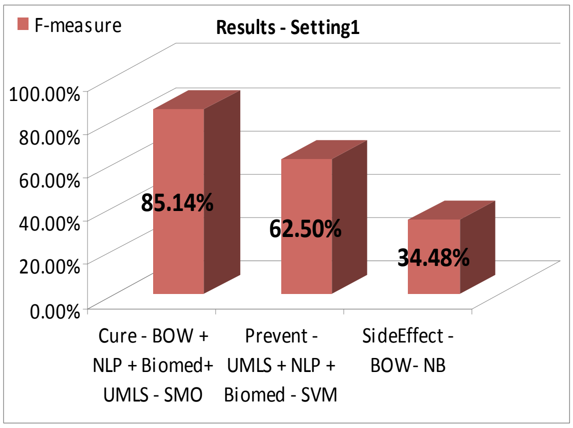
Experiments for the task tackled in our research are performed with all the above-mentioned representations, plus combinations of them. We combine the BOW, UMLS and NLP and biomedical concepts by putting all features together to represent an instance.

5 Results

This section presents the results obtained for the task of identifying semantic relations with the methods described above. As evaluation measures we report F-measure and accuracy values. The main evaluation metric that we consider is the F-measure[9], since it is a suitable when the data set is imbalanced. We report the accuracy measure as well, because we want to compare our results with previous work. Table A1 from appendix A presents the results that we obtained with our methods. The table contains F-measure scores for all three semantic relations with the three experimental settings proposed for all combinations of representation and classification algorithms. In this section, since we cannot report all the results for all the classification algorithms, we decided to report the classifiers that obtained the lower and upper margin of results for every representation setting. More detailed descriptions for the results are present in appendix A. We consider as baseline a classifier that always predicts the majority class. For the relation *Cure* the F-measure baseline is 66.51%, for *Prevent* and *Side Effect* 0%.

The next three figures present the best results obtained for the three experimental settings.

Figure 3. Best results for Setting 1.



[8] http://mmtx.nlm.nih.gov/

[9] F-measure represents the harmonic mean between precision and recall. Precision represents the percentage of correctly classified sentences while recall represents the percentage of sentences identified as relevant by the classifier.

Figure 4. Best results for Setting 2.

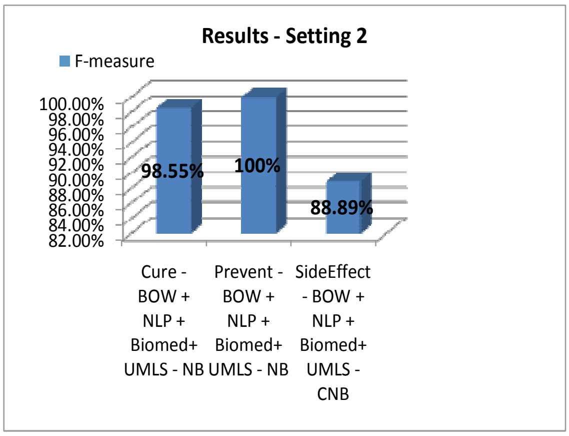
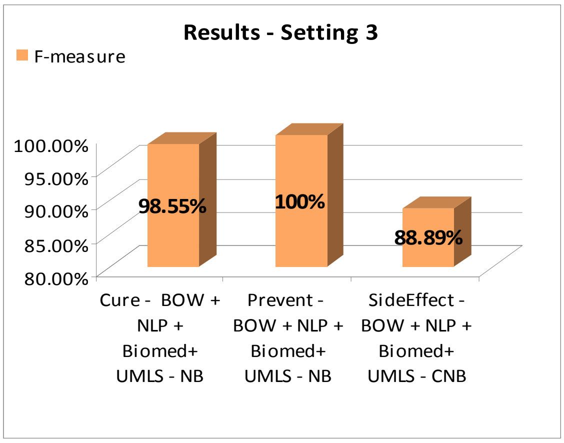


Figure 5. Best results for Setting 3.



6 Discussion

Our goal was to obtain high performance results for the three semantic relations. The first set of experiments was influenced by previous work on a different task. The results obtained show that this setting might not be suitable for the medical domain, due to one of the following possible explanations: the number of examples that are considered as being neutral is not sufficient or not appropriate (the neutral examples are considered sentences that are annotated with a *Vague* relation in the original data); or the negative examples are not appropriate (the negative examples are considered sentences that talk about either treatment or about diseases). The results of these experiments are shown in Figure 3. As future work, we want to run similar setting experiments when considering negative examples sentences that are not informative, labeled *Irrelevant*, from the original data set, and the neutral examples the ones that are considered negative in this current experiments.

In Setting 2, the results are better than in the previous setting, showing that the neutral examples used in the previous experiments confused the algorithms and were not appropriate. These results validate the fact that the previous setting was not the best one for the task.

The best results for the task are obtained with the third setting, when a model is built and trained on a data set that contains all sentences annotated with the three relations. The representation and the classification algorithms were able to make the distinction between the relations and obtained the best results for this task. The results are: 98.55% F-measure for the *Cure* class, 100% F-measure for the *Prevent* class, and 88.89% for the *Side Effect* class.

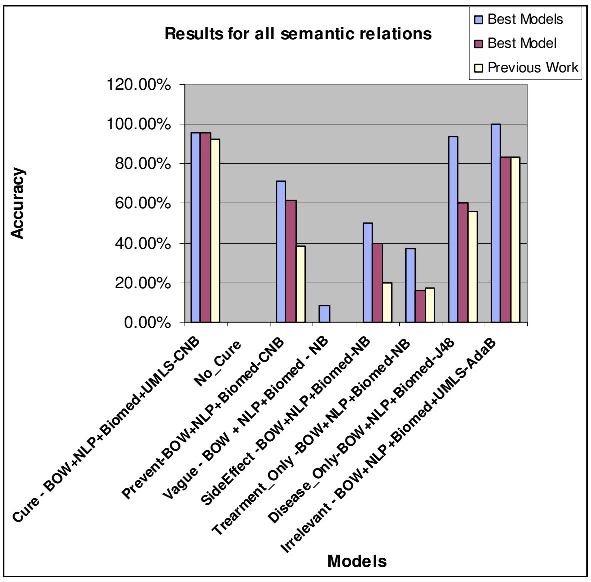
Some important observations can be drawn from the obtained results: probabilistic and linear models combined with informative feature representations bring the best results. They are consistent in outperforming the other classifiers in all the three settings. AdaBoost classifier was outperformed by other classifiers, which is a little surprising, taking into consideration the fact that this classifier tends to work better on imbalanced data. BOW is a representation technique that even though it is simplistic, most of the times it is really hard to outperform. One of the major contributions of this work is the fact that the current experiments show that additional information used in the representation settings brings improvements for the task. The task itself is a knowledge-charged task and the experiments show that classifiers can perform better when richer information (*e.g.* concepts for medical ontologies) is provided.

6.1 Comparison to previous work

Even though our main focus is on the three relations mentioned earlier, in order to validate our methodology, we also performed the 8-class classification task, similar to the one done by Rosario and Hearst (2004). Figure 3 presents a graphical comparison of the results of our methods to the ones obtained in the previous work. We report accuracy values for these experiments, as it was done in the previous work.

In Figure 3, the first set of bar-results represents the best individual results for each relation. The representation technique and classification model that obtains the best results are the ones described on the x-axis.

Figure 3. Comparison of results.



The second series of results represents the overall best model that is reported for each relation. The model reported here is a combination of BOW, verb and noun-phrases, biomedical and UMLS concepts, with a CNB classifier.

The third series of results represent the accuracy results obtained in previous work by Rosario and Hearst (2004). As we can see from the figure, the best individual models have a major improvement over previous results. When a single model is used for all relations, our results improve the previous ones in four relations with the difference varying from: 3 percentage point difference (*Cure*) to 23 percentage point difference (*Prevent*). We obtain the same results for two semantic relations, *No_Cure* and *Vague* and we believe that this is the case due to the fact that these two classes are significantly under-represented compared to the other ones involved in the task. For the *Treatment_Only* relation our results are outperformed with 1.5 percentage points and for the *Irrelevant* relation with 0.1 percentage point, only when we use the same model for all relations.

7 Conclusion and Future Work

We can conclude that additional knowledge and deeper analysis of the task and data in question are required in order to obtain reliable results. Probabilistic models are stable and reliable for the classification of short texts in the medical domain. The representation techniques highly influence the results, common for the ML community, but more informative representations where the ones that consistently obtained the best results.

As future work, we would like to extend the experimental methodology when the first setting is applied, and to use additional sources of information as representation techniques.

References

Ahlers C., Fiszman M., Fushman D., Lang F.-M., Rindflesch T. 2007. *Extracting semantic predications from Medline citations for pharmacogenomics.* Pacific Symposium on Biocomputing, 12:209-220.

Craven M. 1999. *Learning to extract relations from Medline*. AAAI-99 Workshop on Machine Learning for Information Extraction.

Feldman R. Regev Y., Finkelstein-Landau M., Hurvitz E., and Kogan B. 2002. *Mining biomedical literature using information extraction*. Current Drug Discovery.

Friedman C., Kra P., Yu H., Krauthammer M., and Rzhetzky A. 2001. *Genies: a natural-language processing system for the extraction of molecular pathways from journal articles.* Bioinformatics, 17(1).

Ginsberg J., Mohebbi Matthew H., Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant. 2009. *Detecting influenza epidemics using search engine query data.* Nature 457, 1012-1014.

Hunter Lawrence and K. Bretonnel Cohen. 2006. *Biomedical Language Processing: What's Beyond PubMed?* Molecular Cell 21, 589–594.

Ray S. and Craven M. 2001. *Representing sentence structure in Hidden Markov Models for information extraction*. Proceedings of IJCAI-2001.

Rosario B. and Marti A. Hearst. 2004. *Classifying semantic relations in bioscience text.* Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 430.

Koppel M. and J. Schler. 2005. *Using Neutral Examples for Learning Polarity,* Proceedings of IJCAI, Edinburgh, Scotland.

Srinivasan P. and T. Rindflesch 2002. *Exploring text mining from Medline*. Proceedings of the AMIA Symposium.

Appendix A. Detailed Results.

Relation	Representation	Classification Algorithm - F-Measure (%)					
		Setting1		Setting2		Setting3	
Cure	NLP+Biomed	AdaB ZeroR	32.22 66.51	AdaB ZeroR	35.69 67.48	CNB SVM	87.88 94.85
	BOW	AdaB CNB	63.60 79.22	AdaB SVM	67.23 81.43	CNB NB	92.57 96.80
	UMLS	AdaB NB	61.08 74.73	AdaB NB	64.78 76.04	CNB SVM	88.20 95.62
	BOW+UMLS	AdaB CNB	56.07 84.54	AdaB NB	74.68 86.48	J48 NB	96.13 97.50
	NLP+Biomed +UMLS	AdaB NB	61.08 75.18	AdaB NB	64.78 76.70	CNB SVM	90.87 96.58
	NLP+Biomed +BOW	AdaB SVM	53.04 78.98	AdaB CNB	77.46 81.86	J48 NB	96.14 97.86
	NLP+Biomed+ BOW+UMLS	AdaB **SVM**	53.04 **85.14**	AdaB **SVM**	72.32 **87.10**	J48 **NB**	96.32 **98.55**
Prevent	NLP+Biomed	AdaB NB	0 17.02	AdaB,J48 NB	0 22.86	Ada,J48 CNB	0 55.17
	BOW	CNB NB	31.78 50	J48 NB	0 61.9	SVM CNB	50 89.47
	UMLS	AdaB NB	0 28.57	J48 SVM	0 48.28	J48 CNB	0 68.75
	BOW+UMLS	J48 NB	39.02 57.14	J48 NB	9.09 75.68	AdaB CNB	60 89.47
	NLP+Biomed +UMLS	AdaB **SVM**	0 **62.50**	J48 SVM	16 57.69	J48 CNB	0 97.56
	NLP+Biomed +BOW	SVM NB	35 54.90	J48 NB	0 66.67	AdaB CNB	64.52 92.31
	NLP+Biomed+ BOW+UMLS	J48 NB	30.77 62.30	J48 **SVM**	0 **77.78**	AdaB,J48 **NB**	64.52 **100**
Side Effect	NLP+Biomed	AdaB NB,CNB	0 7.69	J48,SVM AdaB	0 18.18	AdaB,J48 CNB	0 33.33
	BOW	AdaB **NB**	0 **34.48**	AdaB,J48 NB	0 50	Ada,J48 CNB	0 66.67
	UMLS	AdaB,J48, SVM NB	0 22.22	J48,SVM NB	0 33.33	AdaB,J48 NB,CNB	0 46.15
	BOW+UMLS	AdaB,J48 NB	0 21.43	J48 NB	0 47	AdaB CNB	0 75
	NLP+Biomed+ UMLS	AdaB,J48 NB	0 19.35	J48 NB	0 31.58	AdaB.J48 NB,CNB	0 46.15
	NLP+Biomed+ BOW	AdaB,J48 NB	0 33.33	J48 **NB**	0 **55.56**	AdaB,J48 **CNB**	0 **88.89**
	NLP+Biomed+ BOW+UMLS	AdaB,J48 NB	0 24	J48 NB	0 46.15	AdaB **CNB**	0 **88.89**

Table A1. Results obtained with our methods.

The *Representation* column describes all the feature representation techniques that we tried. The acronym *NLP* stands from verb and noun-phrase features put together and *Biomed* for bio-medical concepts (the ones extracted by Genia tagger). The first line of results for every representation technique presents the classier that obtained the lowest results, while the second line represents the classifier with the best F-measure score. In bold we mark the best scores for all semantic relations in each of the three settings.