Automatic Recognition of Topic-Classified Relations between Prostate Cancer and Genes from Medline Abstracts

Hong-Woo Chun[1] **Yoshimasa Tsuruoka**[2] **Jin-Dong Kim**[1]
1. Department of Computer Science, Graduate School of
Information Science and Technology, The University of Tokyo
2. School of Informatics, University of Manchester
{chun,tsuruoka,jdkim}@is.s.u-tokyo.ac.jp

Rie Shiba[3] **Naoki Nagata**[4] **Teruyoshi Hishiki**[4]
3. Japan Biological Information Research Center,
Japan Biological Informatics Consortium
4. Biological Information Research Center,
National Institute of Advanced Industrial Science and Technology
{rshiba,nnagata,t-hishiki}@jbirc.aist.go.jp

Jun'ichi Tsujii[1,2,5]
5. SORST, Japan Science and Technology Corporation
tsujii@is.s.u-tokyo.ac.jp

Abstract

To recognize instances of medical information concerning prostate cancer and its relevant genes, we developed a machine learning-based relation recognizer using rich contextual features. We collected prostate cancer-related abstracts from Medline. We then constructed an annotated corpus of prostate cancer and gene relations, which consisted of six $topic-classified$ categories, with more detailed information describing the type of prostate cancer and gene relation. The corpus was made with the help of biologists and a disease and gene dictionary-based name recognition technique. The process of dictionary-based name recognition generates disease-gene pairs that become $candidates$ for biomedically related pairs. Since dictionary matching tends to over-generate candidates, we used a machine learning-based named entity recognition method (1) to provide a feature for each candidate, and (2) to filter out over-generated candidates.

Experimental results showed that using a maximum entropy-based relation recognition method and a maximum entropy-based named entity recognition method together greatly improves precision at the cost of a small reduction in recall for the topic-classified relations.

1 Introduction

Bioinformatics is the application of computer technology to the management and analysis of biomedical data, in which computers are extensively used to gather, store, analyze, and merge it. Many natural language processing techniques are used to extract and analyze useful information from biomedical texts. Examples include recognition of biomedical named entities, such as genes, proteins, cells, tissues and diseases, and extraction of their interactions.

One of the most practical research topics of bioinformatics deals with certain diseases and their relevant genes or proteins. Such information can help researchers such as medical doctors, pharmacists, and biologists to do their work, including diagnosis of disease and development of medicines.

Rosario and Hearst attempted to classify seven semantic relations between the entities $disease$ and $treatment$ using several machine learning techniques which are $hidden\ Markov\ models$ and $neural\ networks$. The seven semantic relations were $cure$, $only\ disease$, $only\ treatment$, $prevent$, $vague$, $side\ effect$ and $no\ cure$. They used the following features: the word itself, its part-of-speech from the Brill tagger, the phrase from which the word was extracted, the word's

MeSH ID, a tri-valued attribute indicating whether the word is a disease or a treatment or neither (based on MeSH), and the word's orthographic characteristics (such as presence of capital letters or numerical digits). There were 1,724 relevant sentences and 3,495 irrelevant sentences, which do not contain both treatment and disease, used for training and testing. Using dynamic hidden Markov models, the authors achieved an F-measure of 0.71. Their results and data show that the most important features for relation classification are the word itself and its MeSH-based attributes (B. Rosario and M. Hearst, 2004).

Chen et al. proposed a method for collecting Alzheimer's disease-related proteins. There were 65 seed genes collected from the OMIM database and mapped to 70 Alzheimer's disease-related proteins in HUGO and SwissProt databases. They then show ranked 765 proteins which were collected using protein interactions of the online predicted human interaction database (OPHID) (Chen et al., 2006).

Chun et al. attempted to extract disease-gene relations using dictionaries and a named entity filtering technique. Their results show that maximum entropy-based named entity filtering improves the performance of disease-gene relations recognition. However, there are two disadvantages using this technique: only 1,000 co-occurrences (sentences) are used for training and testing, and only one kind of relation is considered (Chun et al., 2006).

We aim to recognize relations between prostate cancer and its relevant genes from $Medline$ abstracts. Most biomedical information extraction approaches commonly collect only two biomedical names as a binary relation in accordance with their close proximity. Thus, we defined our $relation$ with six points of view for sentences that contain prostate cancer and gene pairs. We call this approach $topic-classified\ relation\ recognition$. Moreover, numerous previous studies have identified entities or relations that are not grounded in any explicit external model of the world, but rather simply point to substrings in the input text. Such outputs are of intrinsically limited value. For example, a system that produces a table of protein-protein interactions is potentially highly valuable if it refers to specific entities in public databases, but of much more limited value if it lists only potentially ambiguous symbols and names (PSB, 2006). Thus, the output of our research includes the ID tags that are used in six publicly available biological databases: LocusLink, HUGO, SwissProt, RefSeq, DDBJ and UMLS.

2 Topic-classified relation recognition

Figure 1 is an overview of topic-classified relation recognition. Our system first collects sentences that contain at least one pair of disease and gene names, using the dictionary-based longest matching technique. We used a machine learning-based named entity recognition method to provide a feature for each candidate in a machine learning-based topic-classified relation recognition method, and to filter out numerous false positives. We also used a machine learning-based relation recognition method to filter out a lot of false positives.

We have three types of false positives in the dictionary-based results:

- False gene names
- False disease names
- False relations

The results of system then are the topic-classified relations.

2.1 Construction of the gene and disease dictionaries

To link each output entry to publicly available biomedical databases, we created a human gene dictionary and a disease dictionary by merging the entries of multiple public biomedical databases. These dictionaries provide gene- and disease-related terms and cross-references between the original databases.

2.1.1 The gene dictionary

A unique $LocusLink$ identifier for genetic loci is assigned to each entry in the gene dictionary, which enabled us to consistently merge gene information dispersed in different databases. Each entry in the merged gene dictionary holds all the relevant literature information associated with a given gene. We used five public databases to build the gene dictionary: $HUGO$, $LocusLink$, $SwissProt$, $RefSeq$, and $DDBJ$(July 2004). Each entry consists of five items: gene name, gene symbol, gene product, chromosomal band, and PubMed ID tags. Based on these criteria,

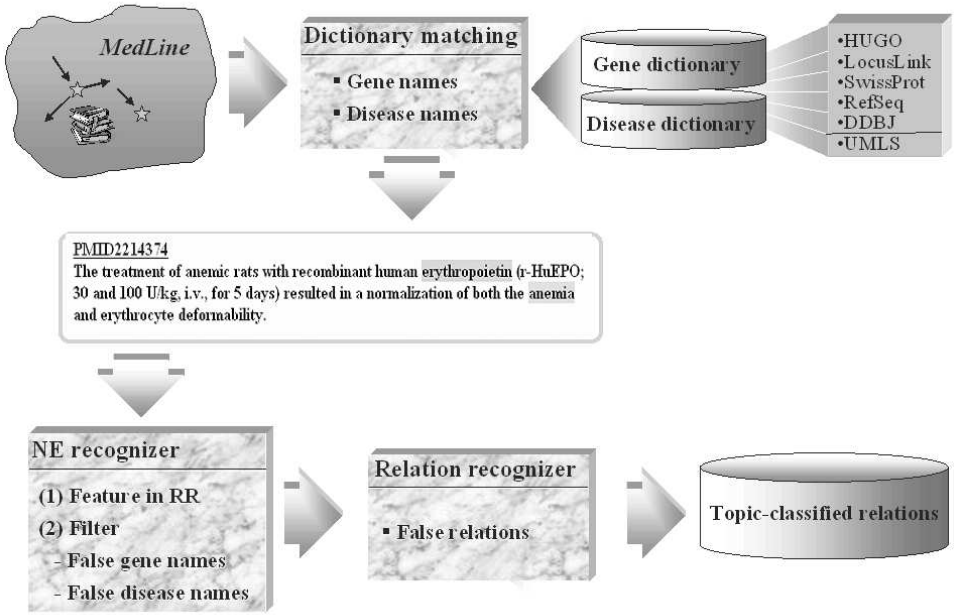

Figure 1: The system architecture.

we created a database-merging system to automatically collect relevant gene information from biomedical data resources. The current version of the gene dictionary contains a total of 34,959 entries with 19,815 HUGO-approved gene symbols, 19,788 HUGO-approved gene names, and 29,470 gene products. It should be noted that there are numerous alias gene symbols and alias gene names in these entries. We found at least 202 approved gene symbols and 253 approved gene names that are used as aliases in different entries or entries without a LocusLink identifier. This tedious merging of data is a result of inconsistencies between databases that cannot be solved simply by combining data into one database. In addition, some words belong to multiple categories and cannot be classified easily into one category. We plan to address these problems in the near future by improving our algorithms. We also want to improve the merging system in order to create other types of dictionaries that will allow comparative genome research.

2.1.2 The disease dictionary

We used the Unified Medical Language System (UMLS) to collect disease-related vocabulary. From the 2003AC edition of the UMLS Metathesaurus, we selected 12 unique identifiers of semantic types (TUIs) that correspond to diseases names, types of abnormal phenomena, or their symptoms (Table 1). From these TUIs, 431,429 unique identifiers for strings (SUIs) for 159,448 unique identifiers for concepts (CUIs) were extracted and stored as a disease-related lexicon.

Table 1: Selected unique identifiers of semantic type (TUIs)

T019	Congenital abnormality
T020	Acquired abnormality
T033	Finding
T037	Injury or poisoning
T046	Pathologic function
T047	Disease or syndrome
T048	Mental or behavioral dysfunction
T049	Cell or molecular dysfunction
T050	Experimental model of disease
T184	Sign or symptom
T190	Anatomical abnormality
T191	Neoplastic process

2.2 Annotation of corpus

The purpose of building an annotated corpus is to construct the training data for machine learning. This annotated corpus has two purposes. It is used as training/testing data to filter out false positives from the dictionary matching results and also to recognize topic-classified relations.

To build training and testing sets, we collected

1,362,285 abstracts through a Medline search using 248 prostate cancer-related terms selected from our disease dictionary. From these abstracts, we generated 2,503,037 co-occurrences [1] using the dictionary-based longest matching technique. Each co-occurrence is a candidate for a topic-classified relation between one prostate cancer and one gene. We chose 3,939 co-occurrences randomly, and they were annotated by four biologists.

For the annotation of prostate cancer-gene relations, we considered three aspects. In other words, the annotator judged a co-occurrence as $correct$ if any of the following three types of relations between the gene and disease had been described in the co-occurrence.

1. Pathophysiology, the mechanisms of diseases, including etiology, the causes of diseases.

2. Therapeutic significance of genes or gene products, more specifically, classification of genes or gene products classified based on their therapeutic use and their potential as therapeutic targets.

3. The use of genes and gene products as markers for disease risk, diagnosis, and prognosis.

2.2.1 Six points of view for sentences that contain prostate cancer and gene pairs

In addition to the $binary\ relation$ between prostate cancer and a gene, we attempted to analyze the following six points of view for sentences that contain prostate cancer and gene pairs.

1. Study description (method)
   In many cases, sentences in the $Methods$ section of papers do not give specific results or conclusions. However, those sentences are still supposed to contain allusive disease-gene relations.

   Example 1 *Thereafter plasma S, cortisol (F) and adrenocorticotropic hormone (ACTH) responses to metyrapone were investigated in 13 normal adult males and 39 patients with prostatic cancer.*

2. Genetic variation
   There are genotypic differences among individuals in a population. For example, mutation (including germ line and somatic), polymorphism (SNP, microsatellite, restriction fragment length), and LOH.

   Example 2 *A polymorphism in endostatin, an angiogenesis inhibitor, predisposes for the development of prostatic adenocarcinoma.*

3. Gene expression
   Gene expression is the phenotypic manifestation of a gene by the processes of genetic transcription and translation. Its profiling is also included.

   Example 3 *The expression of HNK-1 antigen on prostatic cancer was investigated immunohistochemically using the avidin-biotin-peroxidase complex (ABC) method with the anti-HNK-1 monoclonal antibody.*

4. Epigenetics
   Chemical modifications to DNA or histones alter the structure of a chromatin without changing the nucleotide sequence of the DNA.

   Example 4 *Hypermethylation of the 5' promoter region of the glutathione S-transferase pi gene (GSTP1) occurs at a very high frequency in prostate adenocarcinoma.*

5. Pharmacology
   Pharmacology is the science of drugs, including their composition, uses, and effects.

   Example 5 *OBJECTIVES: To assess the involvement of calcitonin gene-related peptide (CGRP) in the occurrence of hot flashes in men after castration for treatment of prostate cancer, we investigated the effects of CGRP on skin temperature in surgically and medically castrated male rats.*

6. Clinical marker
   Measurable and quantifiable gene products are used as biological parameters to indicate health- and physiology-related assessments, such as disease risk, disease diagnosis, cell line development, and epidemiologic studies.

[1] When a sentence contains more than one disease or gene name, the system makes copies of the sentence based on the number of disease-gene pairs. We call these copies $co-occurrences$, they are the input units of our system. For example, if the names of two genes and one disease are referred to in a sentence, then our system makes two co-occurrences from the sentence.

Example 6 *The use of prostate specific antigen (PSA) and digital rectal examination (DRE) results in a three fold increase in prostatic carcinoma detection.*

2.3 Machine learning-based named entity recognition

We used a machine learning-based named entity recognition method for two purposes. One purpose is to provide a feature for each candidate in a machine learning-based topic-classified relation recognition method, and the other is to filter out numerous false positives. The performance of disease and gene name recognition by dictionary matching tends to over-recognize yielding a lot of false positives. A machine learning-based filtering technique can be used to improve the recognition. Maximum entropy models have been used to train the named entity filter. They exhibited the good performances in the CoNLL-2003 shared task of biomedical named entity recognition, and they are widely used in solving classification problems.

2.3.1 Features for named entity recognition

The feature sets used in our experiments were as follows:

- Bag of words:
  All contextual terms in a co-occurrence were considered as a feature.

- Candidate names and contextual terms:
  The features we considered were the candidate name itself by dictionary matching technique as well as unigrams and bigrams. A unigram refers to the word either before or after the candidate name; a bigram refers to the two adjacent words either before or after the candidate name.

- Use of capital letters and numerical digits in the candidate term:
  Capital letters and numerical digits frequently appear in biomedical terms. We accounted for whether or not candidate names contained capital letters and numerical digits.

- Greek letters in the candidate term:
  Greek letters (e.g., $alpha$, $beta$, and $gamma$.) are strong indicators of biomedical terms. Greek letters appear in their original forms such as α, β, $\Gamma(\gamma)$.

Table 2: Affix feature

Prefix/Suffix	Examples
∼cin	actinomycin
∼mide	Cycloheximide
∼zole	Sulphamethoxazole
∼lipid	Phospholipids
∼rogen	Estrogen
∼vitamin	dihydroxyvitamin
∼blast	erythroblast
∼cyte	thymocyte
∼peptide	neuropeptide
∼ma	hybridoma
∼virus	cytomegalovirus

- Affixes of the candidate term:
  Prefixes and suffixes can be very important cues for terminology identification. We considered 11 suffixes given in Table 2. These affixes are commonly used in biomedical terms.

2.4 Machine learning-based topic-classified relation recognition

Disease and gene name pairs co-occurring in a sentence may have some potential relations. We are especially interested in the biomedically meaningful relations defined in the previous section. We developed maximum entropy-based binary classifiers to determine if each pre-defined relation holds between each disease and gene name pair.

2.4.1 Feature set for topic-classified relation recognition

A maximum entropy-based machine learning technique was applied to the co-occurrences in order to recognize instances of meaningful relations. To achieve this, we took into account rich contextual features. The feature set used in our experiments was as follows:

- Bag of words:
  All contextual terms in a co-occurrence were considered as a feature.

- Candidate disease and gene names and contextual terms:
  The features we considered were the candidate disease and gene names themselves as well as unigrams and bigrams of the disease and gene names. We used two kinds of candidate names according to experiments: one is recognized by biologists, and the other is recognized by our machine learning-based named entity recognition method. A unigram

refers to the word either before or after the candidate disease or gene name; a bigram refers to the two adjacent words either before or after the candidate disease or gene name.

- Sequence of candidate names:
  We accounted for the sequence of a candidate gene name and a candidate disease name in each co-occurrence. In other words, we checked whether or not a candidate gene name appeared earlier than a candidate disease name in each co-occurrence.

3 Experimental results

We conducted five sets of experiments for topic-classified relation recognition. The first set of experiments used only the gene and disease dictionary-based longest matching technique. The second set of experiments used not only the gene and disease dictionary matching technique but also disease and gene name filtering. The next three sets of experiments used the maximum entropy-based machine learning technique for topic-classified relation recognition. The third set of experiments used only the maximum entropy-based topic-classified relation recognition and did not use named entity recognition results. The fourth set of experiments used the maximum entropy-based named entity recognition results as features for topic-classified relation recognition. However, the fifth set of experiments used the maximum entropy-based named entity recognition results for filter. We compared the two approaches in the second, fourth and fifth sets of experiments. One approach, which we call $automatic\ NER$, is to use maximum entropy-based named entity recognition results both on training and on testing procedures. The other, which we call $manual\ NER$, is to use human-generated disease and gene names annotation results both on training and on testing procedures. Table 3 lists the performance of all the experiments. The numbers in the first column are the total number of correct answers that are annotated by biologists for each topic-classified relation. We performed 10-fold cross validation to evaluate the systems.

3.1 Performances using dictionary matching (baseline)

The baseline experiment is very simple: we assumed that all prostate cancer-gene pairs recognized by dictionary-based longest matching have a relationship and hold for the all topic-classified relations. The performance of the baseline experiment is listed in the third column of Table 3.

Table 4: Performance of named entity recognition

	Features							Precision	Relative recall
	1	2	3	4	5	6	7	(%)	(%)
GENE								84.4	100.0
	√							93.5	95.4
		√						95.0	98.3
			√					93.1	93.3
				√				84.4	100.0
					√			84.4	100.0
						√		84.4	100.0
							√	84.4	100.0
	√	√	√					94.4	96.1
		√	√					95.0	97.1
		√	√	√				**95.8**	97.0
		√	√		√			94.9	96.7
		√	√			√		94.9	97.0
		√	√	√		√		95.8	96.9
DISEASE								99.2	100.0
	√							99.3	99.8
		√						99.3	100.0
			√					99.3	100.0
				√				99.2	100.0
					√			99.2	100.0
						√		99.2	100.0
							√	99.2	100.0
		√	√	√		√		99.3	100.0

$Note$: 1) Bag of words (all words in co-occurrence); 2) candidate disease and gene names; 3) contextual terms; 4) presence of capital letters in candidate term; 5) presence of numerical digits in candidate term; 6) presence of Greek letters in candidate term; 7) presence of affixes of candidate term.

It should be noted that our dictionaries do not cover all disease and gene names, and thus, we could not calculate the $absolute$ recall in this experiment. Instead, we used $relative\ recall$ as a performance measure. The relative recall is calculated by assuming that the baseline method performs at 100% of this metric. In this approach, we are interested in how precise our system is at correctly identifying the relations, rather than how often it misses other meaningful relations. Thus, we focused on improving the precision.

3.2 Performances using dictionary matching and a disease and gene name filter

We applied named entity recognition techniques to filter out false positives generated by dictionary matching, and we assumed that all the remained prostate cancer-gene pairs have a relationship and hold for the all topic-classified relations. A maximum entropy-based named entity recognition result was used for filter both on training and

Table 3: Experimental results.

Topic-classified Relations (# of correct answers)	(%)	Co-occurrence w/o NER	NER for filter		RR w/o NER	RR with NER for feature		RR with NER for filter	
			Automatic	Manual		Automatic	Manual	Automatic	Manual
Relation	PRE	81.1	91.8	96.7	91.0	91.5	97.0	**92.1**	97.1
(3196)	REC	100.0	97.0	100.0	95.3	96.1	99.6	96.5	99.6
Study description	PRE	26.7	30.2	31.8	65.9	67.5	70.8	**67.6**	70.6
(1050)	REC	100.0	97.2	100.0	57.6	57.6	63.0	62.9	62.9
Genetic Variation	PRE	7.1	8.1	8.4	**79.6**	78.6	81.9	79.4	83.1
(278)	REC	100.0	98.9	100.0	67.3	67.3	70.1	73.6	73.6
Gene Expression	PRE	27.1	30.8	32.3	71.4	73.0	76.2	**73.5**	76.8
(1067)	REC	100.0	97.4	100.0	62.4	61.4	64.5	63.5	64.9
Epigenetics	PRE	1.3	1.6	1.6	87.9	85.7	85.4	**88.1**	88.1
(53)	REC	100.0	100.0	100.0	54.7	67.9	66.0	69.8	69.8
Pharmacology	PRE	9.1	10.3	10.9	64.4	**69.7**	66.7	63.7	67.2
(360)	REC	100.0	96.1	100.0	44.2	45.3	45.0	44.4	45.3
Clinical marker	PRE	31.5	35.9	37.5	75.3	**77.9**	78.2	76.6	78.3
(1240)	REC	100.0	97.8	100.0	70.1	73.2	74.0	73.6	75.4

$Notes$) Co-occurrence: baseline; NER: machine learning-based disease and gene named entity recognition results; RR: machine learning-based topic-classified relation recognition; Automatic: experiments using maximum entropy-based named entity recognition results; Manual: experiments using human-generated annotation results; PRE: precision; REC: relative recall.

on testing procedures in the automatic NER approach. However, a human-generated disease and gene names annotation result was used for filtering both on training and on testing procedures in the manual NER approach. The performances are listed in the fourth and fifth columns of Table 3. The disease and gene name filtering method improves the precision of all topic-classified relation recognitions at the cost of a small reduction in recall. We used the best combination of features that had been attained empirically for named entity recognition.

- Recognition of gene names:
  Candidate names, contextual terms, and presence of capital letters in the candidate term.

- Recognition of disease names:
  Candidate names, contextual terms, and presence of capital letters or Greek letters in the candidate term.

Table 4 shows the performances of named entity recognition using various combinations of features. The first rows for gene and disease names express the performance using dictionary matching. For the disease names, dictionary matching generated very high performance. Thus, the performance of disease name recognition could slightly improve the precision in this experiment.

The performance of $relation$ in the second experiment (in Table 3) was comparatively high. Manual analysis revealed that most correctly identified disease-gene pairs will safely hold for almost any relation[2].

[2]96.7% of all the 2,494 correctly identified disease-gene pairs had been annotated to hold relation.

3.3 Performances using machine learning-based topic-classified relation recognition

We used machine learning techniques for topic-classified relation recognition. Dictionary matching results are the input of this set of experiments, and the best combination of features was considered.

- Relation and clinical marker:
  Candidate gene and disease names, contextual terms, and sequence of candidate names.

- Study description, genetic variation, gene expression, epigenetics and pharmacology:
  Bag of words, candidate gene and disease names, contextual terms, and sequence of candidate names.

This experiment did not consider named entity recognition results, and the sixth column of Table 3 describes the performance. Although the experiment did not consider the disease and gene named entity recognition results, the precision of the machine learning-based topic-classified relation recognition method was much better than that in the baseline experiment.

3.4 Performances using machine learning-based topic-classified relation recognition and named entity recognition results as features

We used the disease and gene names recognition results as features in addition to the contextual features that we considered in section 3.3. A maximum entropy-based named entity recognition result was used as a feature both on training and

Table 5: Performance of NER

Gene	
Precision	95.8%
Relative recall	97.0%
Prostate cancer	
Precision	99.3%
Relative recall	100.0%

on testing procedures in the automatic NER approach (i.e., the seventh column of Table 3). However, a human-generated disease and gene names annotation result was used as a feature both on training and on testing procedures in the manual NER approach (i.e., the eighth column of Table 3). Experimental results showed that using named entity recognition results as features for topic-classified relation recognition improves the performance. We can infer that the disease and gene named entity recognition information is a cogent feature.

3.5 Performances using machine learning-based topic-classified relation recognition and named entity recognition as filter

Named entity recognition results were used to filter out over-generated disease-gene pairs by dictionary matching. A maximum entropy-based named entity recognition result and a human-generated disease and gene names annotation result were used for filtering the over-generated disease-gene pairs both on training and on testing procedures in the automatic NER and manual NER approaches, respectively. Topic-classified relation recognition modules were given only co-occurrences that remained after filtering, and we also used only co-occurrences that remained after filtering to evaluate the performances of the fifth experiment. The performances of these experiments are shown in the ninth and tenth columns of Table 3. We used the same combination of features as those for the experiments in section 3.3. Filtering with named entity recognition results provided higher performance of topic-classified relation recognition than using named entity recognition results as features for machine learning-based topic-classified relation recognition.

4 Conclusions

We have developed machine learning-based topic-classified relation recognizers. Six points of view were used to analyze sentences that contain prostate cancer and gene pairs. Selected Medline abstracts were annotated for these six points of view. A simple dictionary-based longest matching method was tested, which produced numerous false positive results. The annotated abstracts were then input to a maximum entropy-based machine learning module in order to train named entity recognizers and relation recognizers. A comprehensive series of experiments revealed that the machine learning-based approach that used rich contextual features had the potential to improve the performance of topic-classified relation recognition. The effect of two approaches by combining two recognizers was also investigated. The results are encouraging and we are planning several extensions that include incorporating disambiguation techniques (Gaudan et al., 2005) and deep syntactic parsing techniques (ENJU, Tsujii group, 2006) and (Ninomiya et al., 2005). Both classes of techniques have been applied successfully to several tasks, and we expect that incorporating such techniques will supplement our methods by providing appropriate treatment to polysemous terms and richer features of deep syntactic structure.

References

B. Rosario and M. Hearst. 2004. *Classifying Semantic Relations in Bioscience Texts.* Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL).

J. Y. Chen, C. Shen, and A. Y. Sivachenko. 2006. *Mining Alzheimer disease relevant proteins from integrated protein interactome data.* The Pacific Symposium on Biocomputing (PSB) pp. 367–378.

H. Chun, Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki and J. Tsujii. 2006. *Extraction of gene-disease relations from Medline using domain dictionaries and machine learning.* The Pacific Symposium on Biocomputing (PSB) pp. 4–15.

The Pacific Symposium on Biocomputing. 2006. *Linking Biomedical Information Through Text Mining: Session Introduction.* http://helix-web.stanford.edu/psb06/intro-nlp.pdf

S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. 2005. *Resolving abbreviations to their senses in Medline.* Journal of Bioinformatics, 21(18), pp. 3658–3664.

Tsujii group, The University of Tokyo. 2004. *Enju Version 2.1:* http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.html

T. Ninomiya, Y. Tsuruoka, Y. Miyao, and J. Tsujii. 2005. *Efficacy of Beam Thresholding, Unification Filtering and Hybrid Parsing in Probabilistic HPSG Parsing.* Proceedings of the 9th International Workshop on Parsing Technologies.