2005－BIO－3（12）
2005／12／22

Disease-Gene Relations Extraction using Domain Dictionaries and Named Entity Recognition Filtering

Hong-woo Chun[1], Yoshimasa Tsuruoka[1,2], and Jun'ichi Tsujii[1,2,3]

1.Tsujii Laboratory, Room 615, 7th Building of Science,
University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033, Japan
2. CREST, Japan Science and Technology agency,
Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
3. School of Informatics, University of Manchester
POBox 88, Sackville St, MANCHESTER M60 1QD, UK
{chun,tsuruoka,tsujii}@is.s.u-tokyo.ac.jp

Abstract. We extracted disease-gene relations from MedLine using disease/gene dictionaries which are constructed from six public DBs. Since dictionary matching produces a large number of false positives, we developed a method of machine learning-based named entity recognition (NER) to filter out false recognitions of disease/gene names.
We found that the performance of relation extraction depends on the performance of NER filtering and that the filtering improves the precision of relation extraction by 26.7% at the cost of a small reduction in recall.

1 Introduction

Our aim is to extract diseases and their relevant genes from *MedLine* abstracts, which we term *relation extraction*. There are some existing systems for relation extraction from biomedical literature. ArrowSmith [1] and BITOLA [2] extract relations between diseases and genes using background knowledge about the chromosomal location of the starting disease as well as the chromosomal location of the candidate genes from resources such as LocusLink, HUGO and OMIM. G2D [3] also extracts relations by *relative score*, which is calculated by co-occurrence information. An appealing feature of these three systems is that all outputs of these systems are terms used in publicly available biomedical data sources, which means these outputs are linked to such databases and can be used by other researchers. However, these approaches have some problems: Their results could conceivably contain a lot of false positives because they yield too many relations that are dependent only on the co-occurrence information; so many of their results may be unreliable.

There are some studies that employ various NLP techniques in order to obtain high-precision. Proux [4] extracted gene-gene interactions using a part-of-speech (POS) tagger, domain-specific corpora, and a shallow parsing technique. Experimental results show 81% precision and 44% recall. Pustejovsky [5] also used predicate patterns which were built by training from a manually-constructed training corpus. Then they analyzed the subject and the object relation for a main verb to extract them as the arguments for a relation. In this approach, they attempted to recognize entity names by shallow parsing and identify semantic type using a domain ontology, and they dealt with acronym problems and anaphora resolution. Experimental results show 90% precision and 59% recall. The advantages of these approaches are that they considered various contextual features using NLP techniques. However, these approaches have a problem in terms of extracting practical and reusable biological knowledge. The outputs only provide information about relations among the "terms" appearing in text. In other words, the entities in the outputs are not explicitly linked to entities in biological databases. If the outputs provide links to explicit knowledge models, then the utility of these outputs will be increased for other researchers.

In this paper, we extract relations by named entity recognition that consists of two steps. The first step uses a dictionary-based longest matching technique. We create dictionaries constructed from public biomedical databases, which enables us to explicitly link extracted relations with the entries in such databases. Since dictionary-based matching produces many false positives, we filter them out by machine learning in the second step.

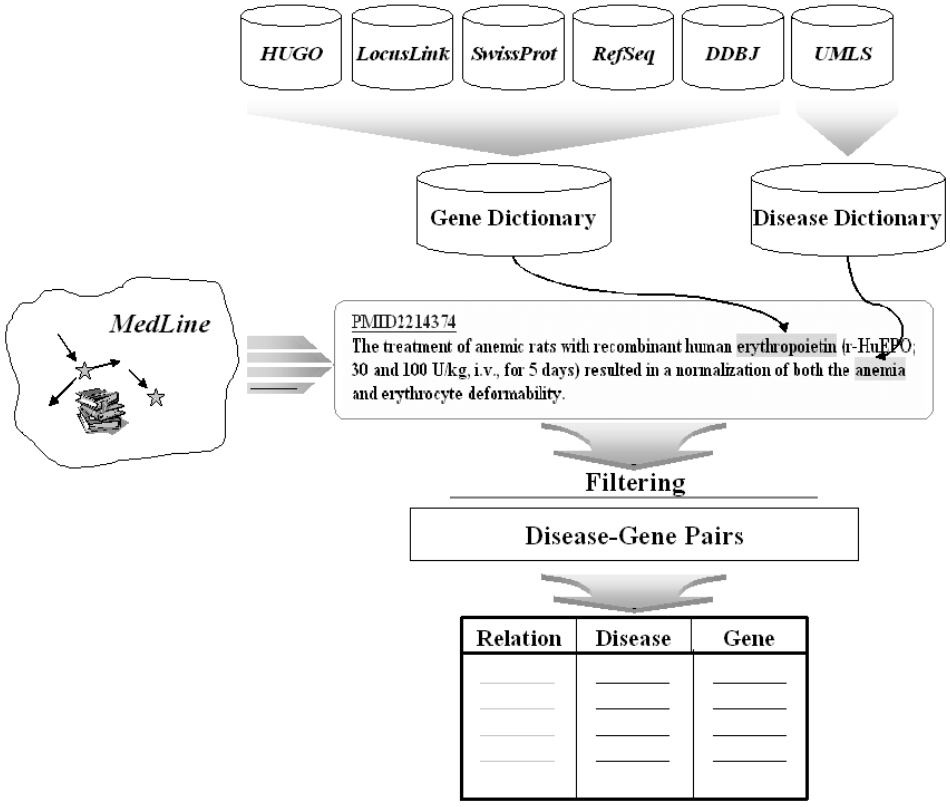

Fig. 1. The system architecture

2 Relation Extraction using Dictionaries and Machine Learning

Figure 1 shows the architecture of our system. Our system first collects sentences that contain at least one pair of disease and gene names, using the dictionary-based longest matching technique. The system then attempts to extract a binary relation between the disease and gene names in each sentence [1].

In this work, we use machine learning to filter out false positives from the dictionary-based longest matching results.

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

- False gene names
- False disease names
- False relations

There are some existing studies in natural language processing aimed at filtering out the first two types of false positives. Tsuruoka and Tsujii [6] proposed a dictionary-based longest matching approach for protein name recognition where they employed a Naive-Bayes classifier to filter out false positives. However, since their dictionary was constructed from the training corpus, their experimental setting is different from the real situation where we have a dictionary constructed from biomedical databases. Furthermore, they used only local context as the features for filtering.

[1] When a sentence contains more than one disease or one gene, the system makes copies of the sentence according to the number of disease-gene pairs. We call each of these copies $co-occurrence$, and regard these items as the input unit of our system. For example, if there are two gene names and one disease name in a sentence, then our system makes two co-occurrences for this sentence.

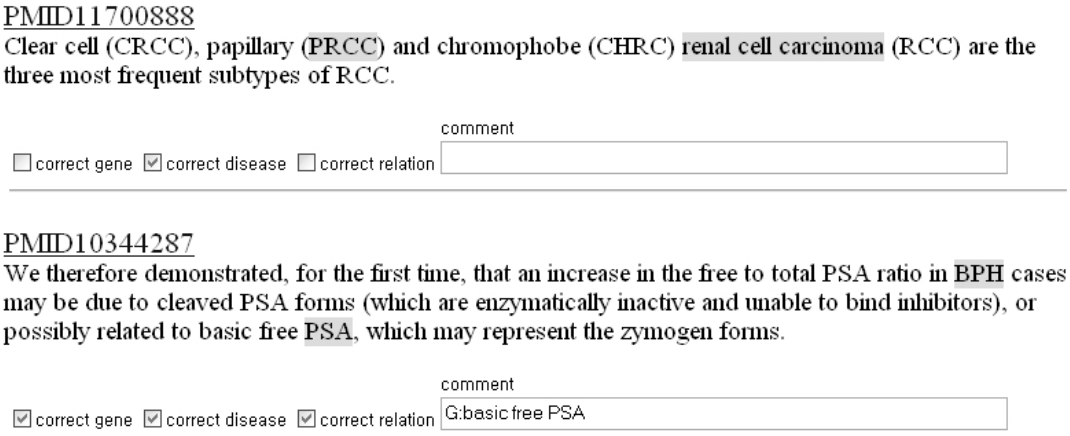

Fig. 2. Example of annotated co-occurrences

2.1 Construction of the Gene and Disease Dictionaries

In order for each output entry to be linked to publicly available biomedical data sources, we created a human gene dictionary and a disease dictionary by merging the entries of multiple public biomedical databases. These two dictionaries provide gene and disease-related terms and cross-references between the original databases.

2.2 Annotation of Corpus

The purpose of building an annotated corpus is to construct the training data for machine learning that will filter out false positives from the dictionary-based results.

To build training and testing sets, 1,362,285 abstracts were collected through a Medline search, using Medical Subject Headings (MeSH) terms. In this work, we used "*Diseases Category*"[*MeSH*] *AND* ("*Amino Acids, Peptides, and Proteins*"[*MeSH*] *OR* "*Genetic Structures*"[*MeSH*]) as the keywords. From the resulting abstracts, we generated 2,503,037 co-occurrences using the dictionary-based longest matching technique. Each co-occurrence is a candidate of a relation between one disease and one gene. We chose 1,000 co-occurrences randomly[2], and they were annotated by one biologist.

Figure 2 shows an example of an annotation. Disease and gene candidates are highlighted: there are four candidates in two co-occurrences. *PRCC* and *PSA* are candidate genes and *renal cell carcinoma* and *BPH* are candidate diseases. These items were recognized by the dictionary-based longest matching technique. The check boxes labeled *correct gene* and *correct disease* are marked by a biologist if he considers the candidates to be correct gene (or disease) names[3].

As for the annotation on disease-gene relations, we considered the following 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 was described in the sentence.

- Pathophysiology, or the mechanisms of diseases, containing etiology, or the causes of diseases.
- Therapeutic significance of the genes or the gene products, more specifically classified to their therapeutic use and their potential as therapeutic targets.
- The use of the genes and the gene products as markers for the disease risk, diagnosis, and prognosis.

[2] We checked all the 1,000 co-occurrences and found that they were all different sentences and they all came from different abstracts.

[3] A name can be embedded in a different name. For example, the dictionary matching may find the disease name *APC* in the term *APC gene*, in which *APC* would be annotated as "incorrect". Embedded names are a major source of false recognitions of gene/disease names.

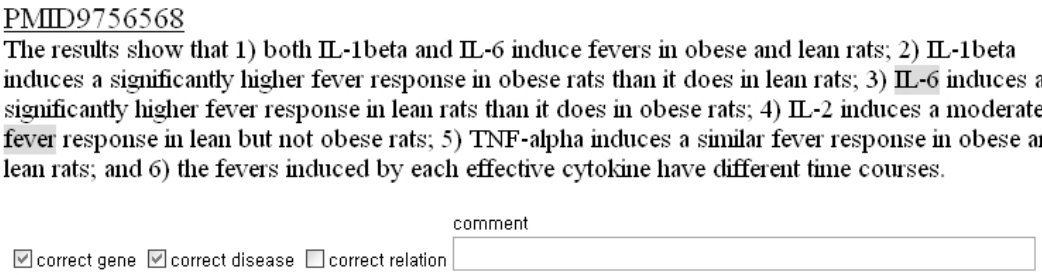

Fig. 3. An example of an annotated co-occurrence whose gene and disease are identified as correct but relation as incorrect

Among 1,000 co-occurrences, 572 co-occurrences contained correctly identified diseases and genes by a biologist. The important observation was that 94% of the 572 co-occurrences were annotated as correct relations, which means that there are few false positives for relations if the disease and gene names are correct. Therefore, we did not perform filtering for relations in this work. Figure 3 shows an example of the remaining 6% of the 572 co-occurrences whose gene and disease were identified as correct but whose relation was incorrect.

2.3 Filtering with a Maximum Entropy-based NER Classifier

To improve the precision of recognizing gene and disease names, we propose the use of a maximum entropy model to filter out false positives. For smoothing, we used Gaussian prior modeling and tuned this parameter with empirical experiments and set it to 300 for genes and 400 for diseases.

Features for NER The feature sets used in our experiments are as follows:

- Candidate names and contextual terms:
  The features we considered were the candidate name itself 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.
- Head word information and the predicate:
  We used the head word information (the word itself and its part-of-speech) of the maximal projection of the disease/gene name as a feature. This analysis is given by the deep-syntactic parser ENJU[4]. In addition, we expect that an important clue for NER is whether or not the candidate is used as an argument of a verb. This is because certain verbs in biomedical literature occur frequently and have a relationship with a disease/gene name; for example, *induce*, *activate*, *contain*, and *phosphorylate*. We named this kind of verb the predicate and considered it as a feature.
- The expanded form of an acronym:
  One of the difficulties in term recognition from biomedical literature is the problem of *ambiguous acronyms*. One acronym can be used with different meanings. We can solve this problem to map the acronym of a candidate name to its full form by scanning the entire abstract. In practice, an acronym and its full form usually occur simultaneously as $full\ form$ ($acronym$) when they first appear in a document.
- Part-of-speech (POS) tags:
  We considered the POSs of the candidate name and its surrounding words. To tag the words with POS labels, we used the *Genia Part-of-Speech Tagger*[5] which is trained on a combined set of the newswire corpus (Penn Treebank) and biological corpus (GENIA corpus[6]).

[4] ENJU v1.0 (2004):
http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.html
[5] GENIA Part-of-Speech Tagger v0.3 (2004):
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/postagger/
[6] GENIA Corpus 3.0p (2003):
http://www-tsujii.is.s.u-tokyo.ac.jp/genia/topics/Corpus/3.0/GENIA3.0p.intro.html

Table 1. 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

– Use of capitals and digits in the candidate term:
Capital characters and numbers frequently appear in biomedical terms. We considered whether candidate names contain capital characters and digits or not.
– Greek letters in the candidate term:
Greek letters (e.g. *alpha*, *beta*, *gamma*, etc.) are strong indicators of biomedical terms. These Greek letters appear in their original forms such as α, β, $\Gamma(\gamma)$.
– Affixes of the candidate term:
Prefixes and suffixes can be very important cues for terminology identification. We considered the 11 suffixes given in Table 1. These affixes are commonly used in biomedical terms.

3 Experimental Results

We conducted two sets of experiments for disease-gene relation extraction. One is an experiment without NER filtering and the other is an experiment with NER filtering.

3.1 Experiments without Filtering (Baseline)

Our baseline experiment is very simple: we assume that all disease-gene pairs recognized by dictionary matching indicate relations. The performance of this baseline experiment is shown in the first row of Table 2.

It should be noted that our dictionaries do not cover all disease/gene names, and thus we cannot calculate the *absolute* recall in this experiment. Instead, we use *relative recall* as a performance measure, and the relative recall given by the baseline method is 100% by definition. In this approach, our interest is in how precise our system is at correctly identifying the relations, rather than how often it misses other meaningful relations.

3.2 Experiments with Filtering

The second set of experiments made use of the maximum entropy-based NER filter. Table 2 lists the performance percentages of relation extraction. We found that NER filtering improves the precision of relation extraction by 26.7% at the cost of a small reduction in recall. This suggests that the performance of relation extraction is very much dependent upon the performance of NER. In this experiment, we used the best combination of features for NER (see Table 3):

– Recognition of Gene names:
Contextual terms, capitalization, Greek letters, POS of disease/gene names and its head, words of predicate and head and full forms if candidate names are acronyms.

Table 2. Relation extraction performance

	Precision(%)	Relative recall(%)
without filtering	51.8	100.0
with filtering	78.5	87.1

Table 3. NER performance

	Features											Precision	Relative recall
	1	2	3	4	5	6	7	8	9	10	11	(%)	(%)
	√	√										86.4	90.2
	√	√	√									85.9	90.2
G	√	√		√								86.2	90.6
E	√	√			√							86.0	90.2
N	√	√				√						86.3	89.4
E	√	√					√					85.9	90.2
	√	√		√		√		√				86.2	90.9
	√	√		√		√				√		86.5	90.5
	√	√		√		√		√	√	√	√	**89.0**	90.9
	√	√										88.5	97.8
D	√		√									88.5	97.9
I	√			√								88.6	98.1
S	√				√							88.6	98.1
E	√					√						88.5	96.0
A	√						√					89.8	95.5
S	√	√				√	√	√				**90.0**	96.6
E	√	√				√	√	√	√	√		89.6	96.6
	√	√				√	√	√			√	89.6	96.0

Note : 1) Candidate disease/gene names and Contextual terms; 2) Use of capitals in the candidate term; 3) Use of digits in the candidate term; 4) Greek letters in the candidate term; 5) Affixes of the candidate term; 6) POS of disease/gene names; 7) POS of disease/gene names and unigram; 8) Head word; 9) POS of head word; 10) Predicates of a candidate disease/gene name; 11) Expanded forms if candidate disease/gene names are acronyms.

– Recognition of Disease names:
  Contextual terms, capitalization, POS of disease/gene names and unigram words and words of head.

All the experimental results for NER considered *contextual terms*. This is because this feature is the most powerful in recognizing candidate names. It leads to improved NER performance of 6.6% for genes and 2.1% for diseases.

4 Conclusion and Future work

The aim of this research was to build a system to automatically extract useful information from publicly available biomedical data sources. In particular, our focus was on relation extraction between diseases and genes. We found that named-entity recognition (NER) using ME-based filtering significantly improves the precision of relation extraction at the cost of a small reduction in recall.

We conducted experiments to show the performance of our relation extraction system and how it depends on the performance of the NER scheme. We could safely regard co-occurrences as containing correct relations if candidate disease and gene names were considered to be correct.

In this work, we did not address the problem of polysemous terms, which would cause difficulty in linking such terms with database entries. One solution would be to incorporate techniques for ambiguity resolution into our system. For example, S. Gaudan et al. [7] proposed the use of SVMs for abbreviation resolution and achieved 98.9% precision and 98.2% recall.

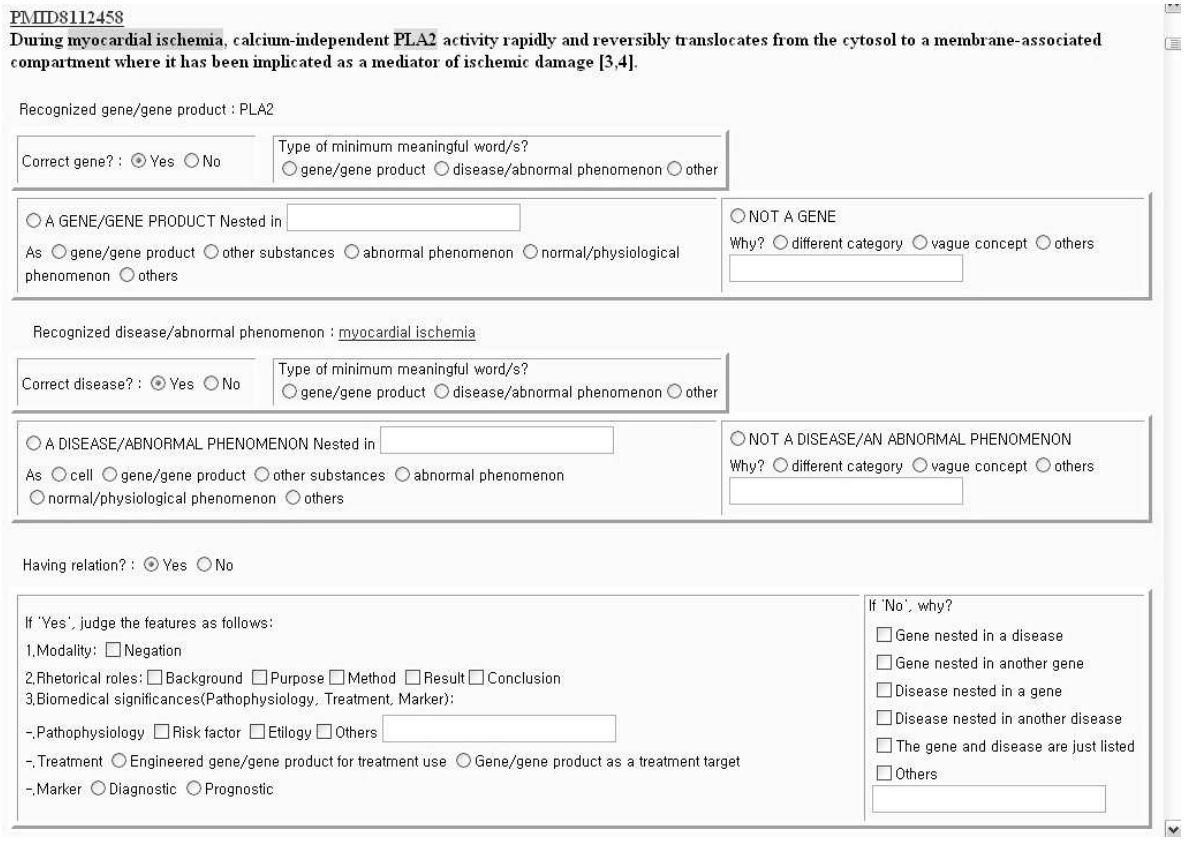

Fig. 4. Sample co-occurrence of annotated corpus

Figure 4 shows the new version of the annotated corpus, which contains more detailed information.

References

1. D.R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspect Biol Med*, 30(1), pp.7–18 (1986).
2. D. Hristovski, B. Peterlin, J.A. Mitchell, and S.M. Humphrey, Improving literature based discovery support by genetic knowledge integration, *Stud. Health Technol. Inform.*, 95, pp.68–73 (2003).
3. C. Perez-Iratxeta, P. Bork, M.A. Andrade, Association of genes to genetically inherited diseases using data mining, *Nat Genet*, 31(3), pp.316–319 (2002).
4. D. Proux et al., A pragmatic information extraction strategy for gathering data on genetic interactions, *ISMB*, 8, pp.279–285 (2000).
5. J. Pustejovsky et al., Medstract : Creating Large-scale Information Servers for biomedical libraries, *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*, pp.85–92 (2002).
6. Y. Tsuruoka and J. Tsujii, Boosting Precision and Recall of Dictionary-Based Protein Name Recognition, *Proc. of the ACL-03 Workshop on Natural Language Processing in Biomedicine*, pp.41–48 (2003).
7. S. Gaudan et al., Resolving abbreviations to their senses in Medline, *Bioinformatics*, 21(18), pp.3658–3664 (2005).