EXTRACTION OF GENE-DISEASE RELATIONS FROM MEDLINE USING DOMAIN DICTIONARIES AND MACHINE LEARNING

HONG-WOO CHUN[1], YOSHIMASA TSURUOKA[1,2], JIN-DONG KIM[1,2], RIE SHIBA[3,4], NAOKI NAGATA[3], TERUYOSHI HISHIKI[3], AND JUN'ICHI TSUJII[1,2,5]

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. Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, AIST Waterfront Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo, 135-0064, Japan
4. Integrated Database Team, Japan Biological Information Research Center, Japan Biological Informatics Consortium, AIST Waterfront Bio-IT Research Building, Aomi 2-42, Koto-ku, Tokyo, 135-0064, Japan
5. School of Informatics, University of Manchester POBox 88, Sackville St, MANCHESTER M60 1QD, UK
E-mail: {chun,tsuruoka,jdkim,tsujii}@is.s.u-tokyo.ac.jp, {rshiba,nnagata,t-hishiki}@jbirc.aist.go.jp

We describe a system that extracts disease-gene relations from *MedLine*. We constructed a dictionary for disease and gene names from six public databases and extracted relation candidates by dictionary matching. 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 is heavily dependent upon 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

The continuing rapid development of the internet makes it very easy to quickly access large amounts of data online. However, it is impossible for a single human to read and comprehend a significant fraction of the available information, and there is a real need for the application of natural language processing techniques in many domains that would facilitate quick and easy

retrieval of useful information. Genomics is not an exception. Databases such as *MedLine* have a vast amount of knowledge.

Our aim in this paper 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 (Swanson 1986) [1] and BITOLA (Hristovski 2003) [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. These systems are designed to discover new, potentially meaningful relations between diseases and genes which do not occur together in the same published article. If concept X and concept Y are related to each other, the systems assume that concepts Z and X have some relationship if Z is relevant to Y. Finally, the systems check whether X and Z appear together in the medical literature. If they do not appear together, this pair (X and Z) is considered as a potentially new relation. G2D (Perez-Iratxeta 2002) [3] also extracts relations by *Relative score*, which is calculated by co-occurrence information. G2D assumes that relevant terms occur together in many abstracts. 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. They have done only a preliminary analysis on the precision of the outputs.

There are some studies that employ various NLP techniques in order to obtain high-precision knowledge from biomedical literature. Proux (2000) [4] extracted gene-gene interactions by manually constructed predicate patterns, which they call scenarios. For example, '[gene product] *acts* as a [modifier] of [gene]' is a scenario of the predicate 'act', which can cover a sentence like: "Egl protein *acts* as a repressor of BicD". In this approach, they employed several techniques for linguistic analysis. Concerning the named entity recognition, they used a part-of-speech (POS) tagger that is based on finite state transducers (FST). This POS tagger contained tokenization and morphological analysis to provide possible POS tags. They used a Hidden Markov Model (HMM) for disambiguation and domain-specific corpora for correcting errors. They then attempted to identify entity names. After that, they did shallow parsing of local structures around verbs to

analyze their subjects and objects and made a conceptual graph using a domain-specific ontology. Experimental results show 81% precision and 44% recall. Pustejovsky (2002) [5] also used predicate patterns. They did not build these patterns manually, but extracted patterns 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.

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 [a].

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

[a] 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.

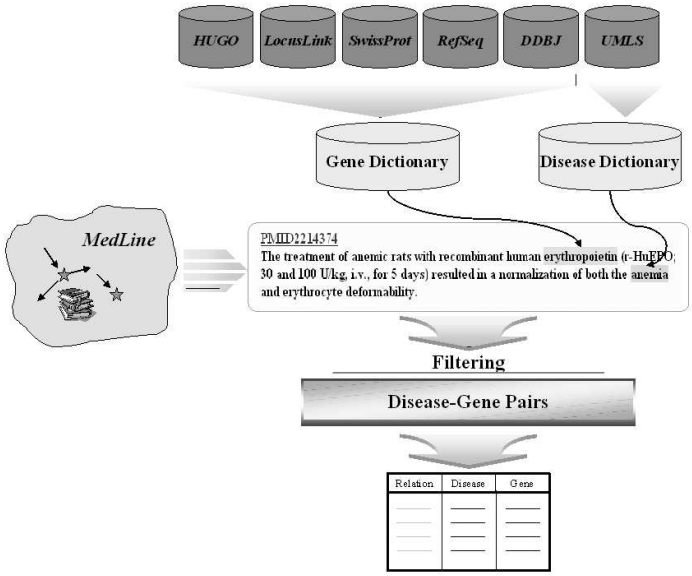

Figure 1. The system architecture

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.

In the following sections, we explain our techniques including dictionaries, a corpus, and the NER filter in detail.

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.

2.1.1. *The gene dictionary*

A unique *LocusLink* identifier for genetic loci is assigned to each entry in the gene dictionary, which enables us to consistently merge gene information dispersed in different databases. Each entry in the merged gene dictionary holds all 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 consisted of five items: gene name, gene symbol, gene product, chromosomal band, and PubMed ID tags. Based on these principles, 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 simply solved by combining data into one database. In addition, some words belong to multiple categories and cannot be easily classified into one category. We plan to address these problems in the near future by improving our algorithms. We also hope to improve the merging system 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 TUIs (unique identifiers of semantic types) that correspond to diseases names, types of abnormal phenomena, or their symptoms (Ta-

Table 1. Selected TUIs (Unique identifiers of semantic type)

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

ble 1). From these TUIs, 431,429 SUIs (unique identifiers for strings) for 159,448 CUIs (unique identifiers for concepts) were extracted and stored as a disease-related lexicon.

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[b], 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

[b]When we checked all 1,000 co-occurrences, we found that they were all different sentences and they all came from different abstracts.

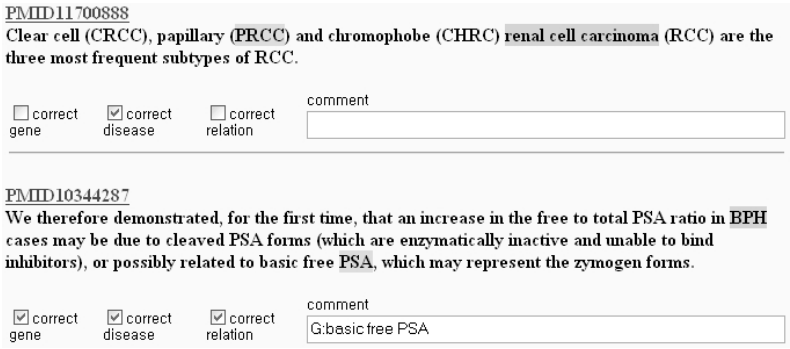

Figure 2. Example of annotated co-occurrences

to be correct gene (or disease) names[c].

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.

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.

[c]Some disease names are embedded in gene names. For example, *APC* is a disease name, but *APC gene* is not a disease name. Therefore, in the case of *APC gene*, *APC* is recognized as a disease candidate by the dictionary-based longest matching technique, but *APC* is not checked as a correct disease name by a biologist.

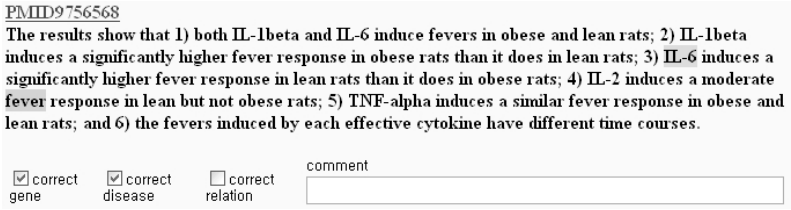

Figure 3. An example of an annotated co-occurrence whose gene and disease are identified as correct but relation as incorrect

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. Maximum entropy models exhibited the best performance in the CoNLL-2003 Shared Task of NER, and are widely used in classification problems in natural language processing. 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.

2.3.1. *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 [7d].

  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 fre-

[d]ENJU achieved 87.85% precision and 86.85% recall for the Penn Treebank and the average parsing time was 360 ms [8].

quently 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 if we have access to its full form. Thus, we tried to map the acronym of a candidate name to its full form by scanning the entire abstract. When coming across an acronym, the system searches for the full form of the acronym and uses the last word of the full form as a feature. 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* [9] which is trained on a combined set of the newswire corpus (Penn Treebank) and biological corpus (GENIA corpus [10]).
- 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 that are $\sim$ *cin*,$\sim$ *mide*, $\sim$ *zole*, $\sim$ *lipid*, $\sim$ *rogen*, $\sim$ *vitamin*, $\sim$ *blast*, $\sim$ *cyte*, $\sim$ *peptide*, $\sim$ *ma*, and $\sim$ *virus*. 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

Table 2. Relation extraction performance

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

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.
- 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.

Table 3. NER performance

	Features											Precision	Relative recall	F-score
	1	2	3	4	5	6	7	8	9	10	11	(%)	(%)	(%)
	√	√										86.4	90.2	88.3
	√	√	√									85.9	90.2	88.0
G	√	√		√								86.2	90.6	88.4
E	√	√			√							86.0	90.2	88.1
N	√	√				√						86.3	89.4	87.8
E	√	√					√					85.9	90.2	88.0
	√	√		√		√		√				86.2	90.9	88.5
	√	√		√		√				√		86.5	90.5	88.4
	√	√		√		√		√	√	√	√	89.0	90.9	**89.9**
D	√	√										88.5	97.8	92.9
I	√		√									88.5	97.9	93.0
S	√			√								88.6	98.1	93.1
E	√				√							88.6	98.1	93.1
A	√					√						88.5	96.0	92.1
S	√						√					89.8	95.5	92.6
E	√	√				√	√	√				90.0	96.6	**93.2**
	√	√				√	√	√	√	√		89.6	96.6	93.0
	√	√				√	√	√			√	89.6	96.0	92.7

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.

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.

One of the ideas to overcome our current problems in merging databases or in recognizing disease/gene names is to tackle the ambiguity problems of abbreviations. There are several other research groups working on the abbreviation problem. S. Gaudan [11] attempted to solve the abbreviation problems using dictionary of abbreviation/sense pairs. They achieved 98.9% precision and 98.2% recall. We plan to incorporate such work into

our system in order to alleviate the problem of ambiguity. The number of co-occurrences in the training and testing sets was rather small for the purpose of evaluating our system. Future work should encompass increasing the size of the annotated corpus and enriching annotation.

References

1. D.R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspect Biol Med*, 30(1):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.*, 95:68–73 (2003).
3. C. Perez-Iratxeta, P. Bork, M.A. Andrade, Association of genes to genetically inherited diseases using data mining, *Nat Genet*, 31(3):316–319 (2002).
4. D. Proux et al., A pragmatic information extraction strategy for gathering data on genetic interactions, *ISMB*, 8:279–285 (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. Enju v1.0: http://www-tsujii.is.s.u-tokyo.ac.jp/enju/index.html (2004).
8. T. Ninomiya, Y. Tsuruoka, Y. Miyao, and J. Tsujii, Efficacy of Beam Thresholding, Unification Filtering and Hybrid Parsing in Probabilistic HPSG Parsing, *Proceedings of the 9th International Workshop on Parsing Technologies* (2005).
9. GENIA Part-of-Speech Tagger v0.3: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/postagger/ (2004).
10. GENIA Corpus 3.0p: http://www-tsujii.is.s.u-tokyo.ac.jp/genia/topics/Corpus/3.0/GENIA3.0p.intro.html (2003).
11. S. Gaudan et al., Resolving abbreviations to their senses in Medline, *Bioinformatics Advance Access published July 21* (2005).