Identifying Gene/Disease Names using Rich Contextual Features for Practical Information Extraction

No Author Given

No Institute Given

Abstract. This paper describes a named entity recognition (NER) approach for gene/disease names which combines dictionary-matching and a machine learning (ML) technique. Our system can provide the ID information about the recognized terms which can be used to link text with biological databases. In this paper, we propose to use a maximum-entropy classifier to incorporate various types of linguistic features. Experimental results show that our approach can significantly improve the precision of recognized terms.

1 Introduction

Easier online access to large text data via the Internet offers many new challenges to researchers trying to automatically extract meaningful information. Genomics is one of the fields that natural language processing (NLP) researchers are particularly interested in because the number of electronic databases is rapidly increasing and a vast amount of knowledge still resides in large collections of biomedical papers such as Medline. In order to extract meaningful information from these data, fundamental studies concerning the recognition of biomedical named entities (NEs), such as genes, proteins, cells, tissues and diseases, are very important.

In this paper, we present a method to recognize gene and disease names in biomedical documents. This work is carried out as a part of the H-invitational project, the purpose of which is to build a human gene database. The final goal of this work is to extract meaningful information from these data that help researchers such as medical doctors, pharmacists and biologists to do their work, including diagnosis of disease, production of medicines.

There are many studies on automatic named entity recognition. In the domain of newspaper texts, satisfactory performance has been achieved [1]. However, named entity recognition in the biomedical domain has not achieved high performance compared with that in the newswire domain. This is mainly due to the following characteristics of biomedical NEs [2]:

- Newly created biomedical terms: In order to express new concepts, many new words and acronyms are created. One can make new words by attaching modifiers before or after NEs, e.g. $activated\ B\ cell\ lines$. Moreover, simplified forms also appear, e.g. $EGFR,\ EGF\ receptor,\ epidermal\ growth\ factor\ receptor$. This kind of extension of biomedical NEs makes named entity recognition (NER) problems very difficult.

- Irregular spelling forms: An entity are written in various spelling forms, e.g. $IL\text{-}2,\ IL2,\ interleukin\ 2, interleukin\text{-}2$.
- Synonyms and acronyms: There can be a lot of synonyms for a term, e.g. Akt is one of PKB's synonyms. Moreover, acronyms are frequently used in the biomedical domain, e.g. $TCED,\ IFN,\ TPA$. In particular, there are a lot of acronyms for gene names.
- Cascading problems: One NE may be embedded in another NE, e.g. $< GENE >< DISEASE > APC < /DISEASE > gene < /GENE >$
- Ambiguity of conjunction and disjunction construction: Two or more NEs share one head noun, e.g. $91\ and\ 84\ kDa\ proteins$. Separating this into two named entities is not straightforward.

These issues often arise in the biomedical literature. We encounter similar problems in our gene and disease name recognition task. In section 2, we describe a dictionary based NER method. In section 3, we introduce a maximum entropy (ME) based NER model and explain the types of features we considered. In section 4, we show experimental results and give some discussion. Finally, we present our conclusions and future work in section 5.

2 Dictionary-based Named Entity Recognition model

Dictionary-based approaches intrinsically provide ID information for the recognized terms which can be used to link a text with biomedical databases. This makes dictionary-based approaches particularly useful as the first step for information extraction from biomedical literature. In order to evaluate the contributions of NLP techniques, we first tried to recognize NEs using a longest matching technique. The number of entries in the dictionary is 44,463 and 159,477 for genes and diseases, respectively. By dictionary-matching, we created 1,000 sentences, each of which contains one pair of gene and disease. We asked two biologists to check all the sentences to evaluate this baseline method. Table 1 shows the results. In this paper, we assume that the dictionaries are comprehensive and the recall of dictionary-matching is 100% (Table 1). The performance is evaluated by using the annotations made by two biologists, so there are two results in Table 1. Table 2 shows the agreement for annotation of NEs made by two biologists. It shows that there is a certain level of disagreement between them.

3 Named Entity Recognition with Maximum Entropy Classifier

To improve the precision of NER, we propose to use maximum entropy (ME) model to filter out false positives. A ME method is a common choice for incorporating various features to tackle classification problems in NLP [3]. Regularization is important in maximum entropy modeling to avoid overfitting to the training data. To avoid this, we used the maximum entropy modeling with inequality constraints [4]. The model gives equally good performance as that of the maximum entropy model with Gaussian

Table 1. The results by dictionary based NER system (%)

	Gene		Disease	
	Result1	Result2	Result1	Result2
Precision	57.7	65.0	78.0	82.1
Recall	100.0	100.0	100.0	100.0
F-score	73.2	78.8	87.6	90.2

Table 2. Agreement for annotation by two biologists (%)

Gene	90.3
Disease	89.3

priors [5]. This model has one adjustable parameter as in Gaussian prior modeling. The parameter is called the $width$ factor. We chose this parameter by empirical experiments to be 300 and 400 for genes and diseases, respectively.

3.1 Feature set

The feature sets used in our experiments are as follows:

- Local words: The features we considered were the NE itself as well as unigrams and bigrams. A unigram refers to the word either before or after the NE; a bigram refers to the two adjacent words either before or after the NE.
- Full parsing results: We considered the full parsing results obtained by analyzing the predicate argument structures in a sentence, which gives a higher level structure for a candidate NE. Of the higher level of structures of a candidate NE, we focused on the head word information (head word, POS of head word) of the maximal projection of the NE. The maximal projection is the highest level of structure. For example, the circle in Figure 1 indicates the maximal projection of a disease candidate $remission$ [6]. In addition, we believe that an important clue for NER and IE is given by the “maximal projection of a candidate NE”; specifically, whether or not it is used as an argument of a verb. This is because special verbs in the biomedical literature frequently occur and have a relationship with a NE; for example, $induce$, $activate$, $contain$, $phosphorylate$.
- Whole abstracts: As is discussed in the introduction, one of the difficulties of NER from biomedical literature is the problem of $ambiguous\ acronyms$. In other words, one acronym can be used with different meanings. If the full form of the acronym is given, the acronym problem may be overcome. Thus, we tried to consider the whole abstract and map the acronym of a NE to its full form. An acronym and its full form usually occur simultaneously when they first appear in a whole document with $full\ form\ (acronym)$. Therefore, we tried to overcome the acronym problem by using the features

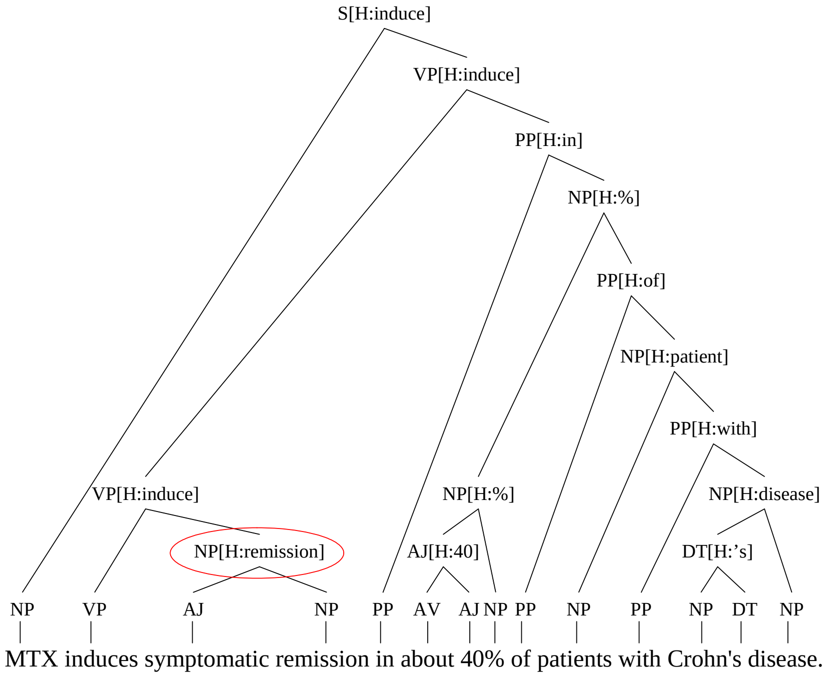


Fig. 1. A maximal projection

extracted from the full form of NEs. We expected that the performance of gene name recognition would be especially improved, because many of the gene names are acronyms.

- POS tags: We considered POS of the NE and its surrounding words. To tag the words with POS labels, we used the $Genia\ Part\text{-}of\text{-}Speech\ Tagger$ which is trained by using a both newswire corpus (Penn Treebank) and biological corpus (GENIA corpus [7]) to achieve high performance on various biomedical documents. It is difficult for general-purpose part-of-speech taggers to achieve high performance on biomedical texts because the character of biomedical texts is quite different from that of newswire articles. We found that the performance achieved by the POS tagger trained on both domains is higher than that achieved by the tagger trained solely on the newswire domain or on the biomedical domain [8].
- Capitalization and digitalization: Capital characters and digit numbers frequently appear in biomedical NEs. Therefore, we consider whether candidate NEs contain capital characters and digit numbers or not.
- Greek letters: Greek letters (e.g. $alpha,\ beta,\ gamma$, etc.) are strongly indicators of biomedical NEs. These Greek letters appear as original forms such as $\alpha,\ \beta,\ \varGamma(\gamma)$.
- Affixes: Prefix and suffix can be a very important cue for terminology identification. We considered the 12 suffixes given in Table 3. These affixes are commonly used in biomedical NEs.

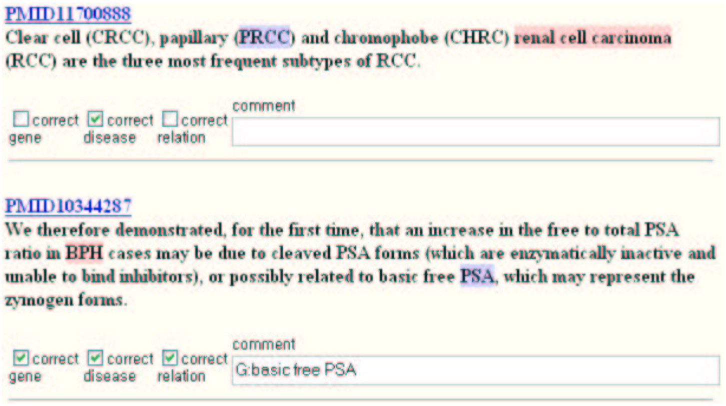
Table 3. Affix feature

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

4 Experiments

4.1 Corpus

We used the same sentences for comparison between a method with NLP techniques and that without NLP techniques. In other words, we selected 1,000 co-occurrences which contain at least one pair of gene and disease names and they were annotated by two biologists. Thus, there are two annotation policies, and we have two annotation results. We constructed experiments with these two annotation results and analyzed how one annotation result can affect the other. Figure 2 shows the interface for annotation. Colored strings indicate the existence of a candidate: blue for gene and pink for disease. They are recognized by the dictionary-based longest matching technique. The check boxes located on the left of *correct gene* and *correct disease* are checked if they are considered as correct NEs by individual biologists. The evaluation of this annotation is shown in Table 1.



PMID11700888

Clear cell (CRCC), papillary (PRCC) and chromophobe (CHRC) renal cell carcinoma (RCC) are the three most frequent subtypes of RCC.

□ correct gene ☑ correct disease □ correct relation

comment

PMID10344287

We therefore demonstrated, for the first time, that an increase in the free to total PSA ratio in BPH cases may be due to cleaved PSA forms (which are enzymatically inactive and unable to bind inhibitors), or possibly related to basic free PSA, which may represent the zymogen forms.

☑ correct gene ☑ correct disease ☑ correct relation

comment: G:basic free PSA

Fig. 2. An example of annotated corpus

4.2 Evaluation

We evaluated our system with 10-fold cross validation. The evaluation measures are *precision*, *recall* and *F-score* [Equation 1, 2, 3].

$$Precision = \frac{\#\ of\ correct\ named\ entities}{\#\ of\ named\ entities\ selected\ by\ the\ system} \tag{1}$$

$$Recall = \frac{\#\ of\ correct\ named\ entities}{\#\ of\ whole\ named\ entities\ in\ a\ given\ corpus} \tag{2}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

In calculating the *F-score*, we gave the same weight to precision and recall. However, considering application fields of this research, we could focus on precision values. In other words, there can be more appropriate measures than those we have presented in these experiments.

4.3 Experimental results

Experiments with different features sets In order to evaluate the contribution of each feature type, we ran experiments using different combinations of features. From the results in Table 4, we can see the following properties.

Table 4. ME based NER (Effects of different features)

	Local	Caps	Digit	Greek	Affix	Part of Speech			Precision	Recall	F-score
	words					NE	NE,Uni	NE,Uni,Bi	(%)	(%)	(%)
	√								82.7	88.3	85.4
	√	√							86.4	90.2	88.3
G	√	√	√						85.9	90.2	88.0
E	√	√		√					86.2	90.6	88.4
N	√	√			√				86.0	90.2	88.1
E	√	√				√			86.3	89.4	87.8
	√	√					√		85.9	90.2	88.0
	√	√						√	85.7	87.5	86.6
	√	√		√			√		**86.7**	89.1	87.9
	√								89.7	95.0	92.3
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	√							√	89.3	95.4	92.2
	√	√					√		**90.0**	96.1	92.9

- Genes
  1. The local words feature is most influential.
  2. When we used the two highest ranked features–local words and capitalization information, our system achieved the highest F-score.
  3. Greek letters play an import role in improving recall values.
  4. When we used local words, capitalization information, Greek letters and POS of NEs, our system achieved the highest precision.
  5. With the features on POS of local words, the use of the POS of unigram words introduces a negative effect on precision, while including bigram words introduces a negative effect on *all* the measures.
  6. Currently, the number of affixes considered is only 12. Our system would show an improved performance by considering more prefixes and suffixes.
- Diseases
  1. The local words feature is again most influential.
  2. Greek letters and affix information significantly improve the recall value, when combined with the local words features.

3. When we used local words, capitalization information, the POS of NEs and unigram words, our system achieved the highest precision.
4. With the POS feature, the inclusion of NE and unigram words is helpful, but including bigrams shows a negative effect.

Experiments with different annotation results Table 5 shows the results obtained by using different annotation results as the data for training and testing. These experiments reveal how much inter-annotator discrepancy affects the performance. For this experiment, we used the feature sets which achieved the highest F-score in the previous section. We used local words, capitalization information and Greek letters for gene recognition and local words and Greek letters for disease recognition. We constructed two additional corpora by combining two annotation results. Namely, when the two biologists agree with a name as an NE, then the Intersection corpus includes it as an NE. When only one biologist considers a name as an NE, then the Union corpus includes it as an NE. Our system achieved the highest precision and the lowest recall when the Intersection corpus is used as the training corpus. On the other hand, our system achieved the highest recall and the lowest precision value when the Union corpus is used as the training corpus.

Table 5. The results according to the training corpus (%)

Test data	Training data	Gene			Disease		
		Precision	Recall	F-score	Precision	Recall	F-score
Annotation1	Annotation1	80.2	83.0	81.6	88.7	95.9	92.2
	Annotation2	77.5	91.5	83.9	86.8	97.6	91.9
	Intersection	81.7	81.3	81.5	89.9	93.3	91.6
	Union	76.5	92.5	83.8	85.5	98.7	91.6
Annotation2	Annotation1	88.6	81.4	84.8	90.4	92.8	91.6
	Annotation2	86.8	90.9	88.8	89.6	95.7	92.6
	Intersection	90.6	80.0	85.0	91.1	89.9	90.5
	Union	85.4	91.7	88.4	88.8	97.4	92.9

Considering the level of agreement of the annotation by the two biologists (Table 2), we can say that the performance of our current system is encouraging.

Experiments considering full parsing results In order to consider not only the information about the NE and its local words but also the information about the structure of the whole sentence, we employed Head-driven Phrase Structure Grammar (HPSG) parsing. We used the $head\ word$ and its POS of the NE as a feature. If there are multiple phrases that include the NE, we chose the head word of the $maximal\ projection$ of NE. We also used $verb$ information as a feature when the maximal projection of the NE is an $argument$ of the verb. The motivation is that there are many biomedical domain specific verbs and they frequently occur with NEs as their argument. We thought that the *verb* feature is an important clue for NER. In this experiment, we did not restrict ourselves to popular verbs in biomedical texts. We considered two cases depending on the relative location of the NE to the verb: one is the case where the NE appears before a verb and the other is the case where the NE appears after a verb. Although this distinction can induce a data sparseness problem, it enables us to analyze the effect of the location of NEs delicately. In this experiment, we consider four types of features: local words, capitalization information, Greek letters and parsing results for gene recognition. In addition, we consider four other types of features: local words, capitalization information, POS of candidate NE and unigram and parsing results by empirical experiments in previous two sections. At this time, we use the second annotation result in 4.3.2. Finally, the head words play an important role in improving performance of NER (Table 6).

Table 6. ME based NER (Using parsing results)

	Head word	PoS (Head)	Verb	Precision (%)	Recall (%)	F-score (%)
	√			86.2	90.9	**88.5**
G		√		85.9	90.8	88.3
E			√	86.5	90.5	88.4
N	√	√		85.8	**91.2**	88.4
E	√		√	86.4	90.6	88.4
		√	√	86.0	90.9	88.4
	√	√	√	85.9	**91.2**	**88.5**
D	√			**90.0**	**96.6**	**93.2**
I		√		89.8	96.3	92.9
S			√	89.9	96.3	93.0
E	√	√		89.5	**96.6**	92.9
A	√		√	**90.0**	96.5	93.1
S		√	√	89.7	96.5	93.0
E	√	√	√	89.6	**96.6**	93.0

Experiments considering whole abstracts In order to alleviate the problem of ambiguous acronyms, we use the whole abstract rather than the sentence. This means we can resolve the ambiguity by using the information about the full form of an acronym. An acronym and its full form often occur simultaneously when they first appear in a biomedical document. Finally, we found that this method can increase the performance of recognition for gene name (88.1% precision, 91.4% recall and 89.7% F-score).

Learning curve Figure 3 and Figure 4 show the effect of the size of the training corpus. In this experiment, we only used the local words features. From these two figures, we can see that the size of the training corpus plays an important role in improving the performance in *all* the measures. Therefore, we can expect that our system will achieve a higher performance if we have a larger training corpus.

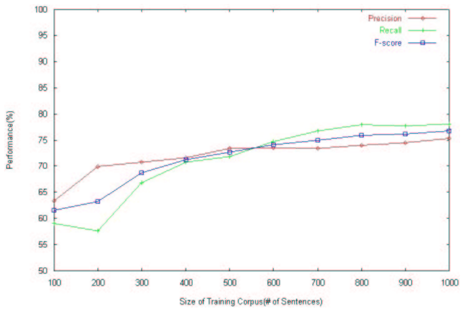


Fig. 3. Gene recognition performance and the size of the training corpus

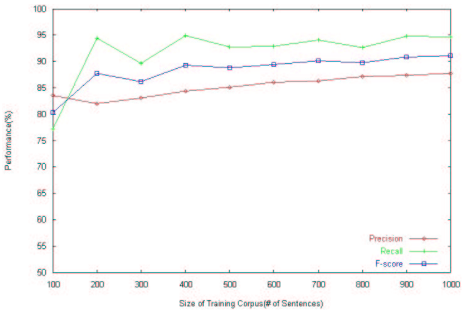


Fig. 4. Disease recognition performance and the size of the training corpus

5 Conclusion

This paper presents the development of a new NER system consisting of two phases: dictionary-based candidate recognition and ML-based filtering. In the dictionary-based phase, we scan texts for gene/disease names using gene/disease dictionaries. In the ML-based phase, we filter the results of the first phase by using a maximum entropy classifier and rich linguistic features. Our method can provide ID information about the recognized terms because we use a dictionary. Therefore, this research is the first step of practical information extraction technique for identifying the interactions between genes and diseases described in the biomedical literature.

Experimental results show that our approach is encouraging. We achieved 88.1% precision, 91.4% recall and 89.7% F-score for genes and 90.0% precision, 96.6% recall and 93.2% F-score for diseases.

Our future work should encompass to explore more influential features and effective methods to cope with NER difficulties. We expect that our system can achieve higher performance by considering more influential features and increasing the size of the training corpus.

Acknowledgements

This research is part of the H-invitational project with $JBIRC$(Japan Biological Information Research Center). This work was partially supported by Grant-in-Aid for

Scientific Research on Priority Areas (C) “Genome Information Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Guodong Zhou and Jian Su, Named Entity Recognition using an HMM-based Chunk Tagger. *Proc. 40th Annual Meeting of the Association for Computational Linguistics(ACL)*, pp.473–480, 2002.
2. Jie Zhang, Dan Shen. Guodong Zhou, Jian Su and Chew-Lim Tan, Exploring Various Evidences for Recognition of Named Entities in Biomedical Domain. *Proc. EMNLP*, 2003.
3. Adam L. Berger, Stephen A. Della Pietra and Vincent J. Della Pietra, A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1): pp.39–71, 1996.
4. Jun’ichi Kazama and Jun’ichi Tsujii, Evaluation and extension of maximum entropy models with inequality constraints. *Proc. EMNLP*, 2003.
5. Stanley F. Chen and Ronald Rosenfeld, A gaussian prior for smoothing maximum entropy models. *Technical Report CMUCS, Carnegie Mellon University*, 1999.
6. Peter Sells, Lectures on contemporary syntactic theories *Center for the study of language and information*, 1985.
7. GENIA Corpus 3.0p:
http://www-tsujii.is.s.u-tokyo.ac.jp/genia/topics/Corpus/3.0/GENIA3.0p.intro.html (2003)
8. GENIA Part-of-Speech Tagger v0.3: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/postagger/ (2004)