Exploiting the contextual cues for bio-entity name recognition in biomedical literature

Zhihao Yang*, Hongfei Lin, Yanpeng Li

Department of Computer Science and Engineering, Dalian University of Technology, No. 2 LingGong Road, ShuHeKou District, Dalian 116023, China

Received 18 May 2007
Available online 11 January 2008

Abstract

To extract biomedical information about bio-entities from the huge amount of biomedical literature, the first key step is recognizing their names in these literatures, which remains a challenging task due to the irregularities and ambiguities in bio-entities nomenclature. The recognition performances of the current popular methods, machine learning techniques, still have much space to be improved. This paper presents a Conditional Random Field-based approach used to recognize the names of bio-entities including gene, protein, cell type, cell line and studies the methods of improving the performance by the exploitation of the contextual cues including bracket pair, heuristic syntax structure and interaction words cue. Experiment results on both JNLPBA2004 and BioCreative2004 task 1A datasets show that these methods can improve Conditional Random Field-based recognition performance by more than 2 points in F-score.

Keywords: Text mining; Information extraction; Named entity recognition; Conditional random fields; Contextual cue

1. Introduction

Along with the rapid expansion of biomedical literature, the demand for efficiently extracting biomedical information from the huge amount of resources offer an excellent opportunity for biomedical text mining, i.e., the automatic discovery of biomedical knowledge. Among others, extracting relationship between bio-entities from biomedical literature has become a research focus. To accomplish it, the fundamental task is named entity recognition (NER), which is the identification of text terms referring to items of interest. In biomedical domain, named entities (called bio-entities) include gene, protein, cell type, cell line, etc. Only when these bio-entities are correctly identified could their relationship be extracted correctly.

NER is not a new task in text mining. In previous research work, many NER systems have been applied successfully in the newswire domain. But in biomedical domain it remains a challenging task due to the irregular-
Fukuda et al. [8] and Olsson et al. [9] proposed rule-based approaches. The former exploited surface clues and parts of speech. The latter, in addition to surface clues, used a syntactic parser for determining protein name boundaries. Olsson et al. conducted experiments for comparing their system (Yapex) with Fukuda’s system (Kex) on 200 MEDLINE abstracts and reported that they achieved a recall of 66.4% and a precision of 67.8% on Yapex and a recall of 41.1% and a precision of 40.4% on Kex in terms of exact matching.

Machine learning techniques have an advantage that they can identify potential bio-entities which are not previously included in standard dictionaries. Currently, there are some research efforts using machine learning techniques to recognize bio-entities in texts. These techniques include HMM [10], SVM [11], MEMM [12], CRFs [13], etc. In JNLPBA2004, Settles achieved an F-score of 69.8% using CRFs with only several kinds of features and no external resource [13].

However, the recognition performances of machine learning techniques such as HMM, MEMM, CRFs depend heavily on the quality and quantity of the training set and the selection of feature set. But it is a time-consuming and costly work to build a large and qualified training set. For example, GENIA corpus is the largest corpus of its type currently available, comprising 2000 abstracts with 18,545 sentences containing 39,373 named entities. However, in GENIA corpus many entities were doubly classified as “protein molecule or region” and “DNA molecule or region”, suggests that inter-annotator agreement could be low, and that many entities in fact have more than one classification.

Another area where GENIA appears inconsistent is in the labeling of preceding adjectives. Take “activated protein molecule or region” for example. Of the 48 times it occurred before or at the beginning of recognition performance by more than 2 points in F-score. The remaining part of this paper is organized as follows: Section 2 describes our method. Section 3 presents and discusses the experiment results on both JNLPBA2004 and BioCreative2004 task 1A datasets. Finally, Section 4 offers some concluding remarks.

2. Methods

Our method is a Conditional Random Field-based method. First we use a CRF model trained on the JNLPBA2004 (or BioCreative2004 task 1A) training set to recognize the bio-entities. Then the results are further processed via the exploitation of the contextual cues including bracket pair, heuristic syntax structure and interaction words cue. The details are described in the following sections.

2.1. CRF recognition

2.1.1. Conditional random fields

Bio-entity recognition can be thought of as a sequence segmentation problem: each word is a token in a sequence to be assigned a label (e.g. protein, RNA, DNA, cell line, cell type, or other). Conditional Random Fields are undirected statistical graphical models, a special case of which is a linear chain that corresponds to a conditionally trained finite-state machine. Such models are well suited to sequence analysis.

Let \( o = (o_1, o_2, \ldots, o_n) \) be a sequence of observed words of length \( n \). Let \( S \) be a set of states in a finite-state machine, each of which is associated with a label \( e \in L \). Let \( s = (s_1, s_2, \ldots, s_n) \) be the sequence of states in \( S \) that correspond to the labels assigned to words in the input sequence \( o \). Linear chain CRFs define the conditional probability of a state sequence given an input sequence to be:

\[
P(s \mid o) = \frac{1}{Z} \exp \left( \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_k f_k(s_{i-1}, s_i, o, i) \right)
\]

where \( Z \) is a normalization factor of all state sequences, \( f_k(s_{i-1}, s_i, o, i) \) is one of \( m \) functions that describes a feature, and \( \lambda_k \) is a learned weight for each such feature function. The training process is to find the weights that maximize the log likelihood of all instances in training data:

\[
LL(D) = \sum_j \log(P(s_j \mid o_j)) - \frac{1}{2} \sum_k \lambda_k^2
\]

The second term in Eq. (2) is a spherical Gaussian prior over feature weights. Once these settings are found, the labeling for a new, unlabeled sequence can be done using a modified Viterbi algorithm. We use first-order leaner chain CRFs and LBFGS training method [14]. CRFs are presented in more complete detail by Lafferty et al. [15].

2.1.2. Feature set

Feature based statistical models like CRFs reduce the problem to finding an appropriate feature set. We used the following features:

...
2.2. Exploitation of contextual cues

Through the analysis of the CRF recognition errors, we found the performance could be further improved via the exploitation of the contextual cues. We exploited three kinds of contextual cues: bracket pair, heuristic syntax structure and interaction words cue.
set. First we extracted all the entity names in the training set and then calculated the frequencies of the last words of these entity names. The last words whose frequency is higher than a certain threshold are added to the Post-keyword list. In this case, “disorders” is not a Post-keyword of any class so “PTLDs” is tagged as “O”. In Table 3 “Toll-like receptor 2” is tagged as DNA class by CRF model, but then is adjusted to protein class according to its Post-keyword “receptor”, which is a high frequent Post-keyword of protein class.

In addition, the newly found entity names in one Medline record are added to a list for later recognition in the following sentences of this record and those found not bio-entity names are also added to a list for filter out the other false positives in this record. Here the reason we didn’t introduce external dictionary is that there exists the annotation ambiguity problem: an entity name may be annotated as one class type in one context and annotated as another class type in another context. For example, “CD28” can refer to a protein or DNA in different contexts. So we confine the effect of these two lists to current Medline record since in most cases an entity name should be annotated as one class type in one record.

2.2.1.2. Non-full name–abbreviation pair. Besides full name–abbreviation pairs, our method can also process non-full name–abbreviation pairs such as “B-protein lymphoid Specific octamer binding protein (OTF-2B)”, which is also a common phenomenon in biomedical literature. In this case, if the long form “B-protein lymphoid Specific octamer binding protein” is recognized as an entity and the short form “OTF-2B” which has the features of proteins (genes) name, e.g. the first letter is uppercase, including digit or dash) will be classified as the same class of the long form.

2.2.2. Heuristic syntax structure

In biomedical literature, there are some heuristic syntax structures that imply the existences of some bio-entities and the classes they belong to. Table 4 shows some heuristic syntax structure examples.

The examples 1–6 demonstrate the syntax structures that help to identify bio-entities and their class. For example, in the example 1, “NFX1.1” and “NFX1.2” can be reasonably recognized as proteins since they are “complexes” and “complexes” is a high frequent Post-keyword of protein class. While the example 7 help to identify that “D609” should be tagged as “O” since it is an “electrolyte” and “electrolyte” is not a Post-keyword of any class. To find the heuristic syntax structures in text we compiled some heuristic syntax structure patterns manually. Most of them are the appositive and copula structures that are indicative to find the is–a relations.

In addition, we found that two similar names (similar length and structure) connected by “and” or “or” often belong to the same class. For example, “NF-YA” and “NF-YB” all belong to the protein class. We call such two similar names Analogy Names. If one of Analogy Names is tagged as one class and the other is tagged “O”, then it is reasonable to tag them as the same class. Furthermore, our method not only uses the coordination or appositive structures expressed by “and” and “or” (which connects two bio-entities) but also those expressed by “,” (which connects more than two bio-entities). For example, in sentence “an immunohistochemical study including correlation with cathepsin D, type IV collagen, laminin, fibronectin, EGFR, c-erbB-2 oncoprotein, p53...”, “cathepsin D”, “type IV collagen”, “laminin”, “fibronectin”, “EGFR”, “c-erbB-2 oncoprotein”, “p53” are coordination structure. If some of them are tagged as one class, it is reasonable to tag the rest of them as the same class.

2.2.3. Interaction words cues

In biomedical literature, the occurrences of some high frequent protein(gene) interaction verbs like “bind”, “interact”, “activate”, “inhibit” and their variants “binding”, “interaction”, “activation”, “inhibition” usually imply the existences of some protein(gene) names nearby (we call them interaction words cues). Table 5 shows some interaction words cue examples.

An interaction word list of about 150 entries (including interaction verbs and their variants) was constructed. When an interaction word is detected in text, its subject and object are checked if they look like proteins (genes)
(e.g. the first letter is uppercase, including digit or a high frequent Post-keyword of protein (gene) class). Here interaction word’s adjacent previous NP and adjacent next NP are used as its subject and object.

3. Experimental results

3.1. Datasets

We conducted experiments using both the JNLPBA2004 and BioCreative2004 task 1A datasets. The JNLPBA2004 training set is GENIA corpus version 3.02 and includes 2000 MEDLINE records retrieved using the MeSH terms “human”, “blood cells” and “transcription factors”. The test set includes 404 MEDLINE records and half of them were from the same domain as the training data and the test set includes 404 MEDLINE records and half of them were from the super-domain of “blood cells” and “transcription factors”. JNLPBA2004 required participating systems to identify the five named entities of protein, RNA, DNA, cell line and cell type. BioCreative 2004 task 1A dataset consists of 20,000 sentences where 15,000 were chosen for training and the other 5000 sentences for evaluation. Entities labeled in dataset have only one type that combines proteins, DNAs, RNAs into one class labeled as “NEWGENE”.

3.2. CRF recognition result

Our CRF recognition result is shown in Table 6. Results are given as F-scores defined as $F = (2PR)/(P + R)$, where $P$ denotes Precision and $R$ Recall. The last line is the result achieved on BioCreative2004 task 1A dataset.

On JNLPBA2004 dataset our CRF model achieved an F-score of 71.87% which is fairly good performance compared with the top JNLPBA2004 systems and only lower than the best system (72.6%) while on BioCreative2004 task 1A dataset it achieved an F-score of 81.32% which is lower than the best system (83.2%).

It need to be pointed out that JNLPB2004 adopted the exact matching criterion: a candidate NE can only be counted as a match if both its boundaries and its class fully coincide with an annotated NE while BioCreative2004 task 1A adopted the relax matching criterion: it allowed several correct annotations, of which NER systems need only match one (e.g. both “no correlation between serum <gene>LH</gene>” and “no correlation between <gene>serum LH</gene>” are correct in BioCreative2).

We made comparison among four CRF-based methods: McDonald and Pereira [18], Tsai et al. [19], Settles [13] and ours (here respectively represented using A, B, C and D for short). All the four have used CRF as the main framework of machine learning model and the configuration is almost the same: first-order leaner chain CRFs and LBFGS training method. Surface Word Features, Orthographic Features, Prefix/Suffix Features, Word Shape Features and feature conjunction are almost the same in the four systems. However, there are some minor differences among the four systems. A and C did not make use of POS and chunking features, which has been used in B and D. From the results comparison, these features improve the performance by around 1 point in F-score. In addition, D proposed two types of feature i.e., keywords feature and boundary term feature which were not used in the other three systems and improved the performance by 0.5 point in F-score. A made use of gene/protein lexicon information to enhance the performance and applied an automatic feature induction method. C has also used dictionary features by constructing a lexicon of five types of entity, but the effect was little, therefore these are not included in their final system. The main original work of B is so called numerical normalization, and pattern-based post-processing which achieved an improvement of F-score is 1.61 points (from 71.37% to 72.98%). Besides, their window size is five, and all others are three. In our recent research, we find that systems with top performances in BioCreative II have used windows with the size of 5 for feature conjunction [20] [21]. The results of these four methods are shown in Table 7. Before post-processing our CRF recognition results are not best. However, after post-processing with the methods of exploiting the contextual cues our final recognition results become best (As described in Section 3.3).

<table>
<thead>
<tr>
<th>Table 6 CRFs recognition results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
</tr>
<tr>
<td>Protein</td>
</tr>
<tr>
<td>DNA</td>
</tr>
<tr>
<td>RNA</td>
</tr>
<tr>
<td>Cell type</td>
</tr>
<tr>
<td>Cell line</td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td>BioCreative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7 CRF recognition performance comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
</tr>
<tr>
<td>A (BioCreative)</td>
</tr>
<tr>
<td>B (JNLPBA)</td>
</tr>
<tr>
<td>C (JNLPBA)</td>
</tr>
<tr>
<td>D (JNLPBA)</td>
</tr>
<tr>
<td>D (BioCreative)</td>
</tr>
</tbody>
</table>

A and C did not have apply post-processing methods so their F-scores before and after post-processing are equal.
The performances of DNA and cell line are relatively inferior mainly due to the small amount of instances of this class in training corpus. However, RNA is an exception because of its regular nomenclature: most RNA instances end with some high frequent Post-keywords such as “mRNA(s)”, “RNA(s)”, and “transcript(s)”. In fact, the nomenclatures of cell type and cell line are also relatively regular: most cell type instances end with “cell(s)”, “lymphocyte(s)”, and “lineage(s)”; most cell line instances end with “cell(s)”, “line(s)”, “lymphocyte(s)”, and “clone(s)”. However, cell line and cell type class are difficult to distinguish even for an expert since many instances of the two classes end with some common words such as “cell(s)” and “lymphocyte(s)”. The Precision of cell line is rather low (57.17%) which shows that more instances of cell type are mistakenly tagged as cell line. In addition, the number of cell line instances in test set is relatively small (500 instances) and its performance is more sensitive to the influence of wrong tagging between cell type and cell line, which is the main reason for its poor performance (61.84% in F-score). In fact, combining cell type and cell line into another single class (we call it “Newcell”) can obtain an F-score of 78.80% which is higher than the ones of cell type and cell line (75.78% and 61.84%).

Table 9 shows the detailed performance on JNLPBA2004 dataset. The last two lines are Right Boundary Correct and Left Boundary Correct performance. The former (81.03% in F-score) is much higher than Fully Correct (75.04% in F-score) which shows that many errors occurred at left boundaries of long names.

Since our method is a CRF-based one, the recognition performance of one class mainly depends on the amount of instances of this class in training corpus. The instance numbers and performances of the different classes are shown in Table 10. The performances of DNA and cell line are relatively inferior mainly due to the small amount of instances of these classes in training corpus. However, RNA is an exception because of its regular nomenclature: most RNA instances end with some high frequent Post-keywords such as “mRNA(s)”, “RNA(s)”, and “transcript(s)”. In fact, the nomenclatures of cell type and cell line are also relatively regular: most cell type instances end with “cell(s)”, “lymphocyte(s)”, and “lineage(s)”; most cell line instances end with “cell(s)”, “line(s)”, “lymphocyte(s)”, and “clone(s)”. However, cell line and cell type class are difficult to distinguish even for an expert since many instances of the two classes end with some common words such as “cell(s)” and “lymphocyte(s)”. The Precision of cell line is rather low (57.17%) which shows that more instances of cell type are mistakenly tagged as cell line. In addition, the number of cell line instances in test set is relatively small (500 instances) and its performance is more sensitive to the influence of wrong tagging between cell type and cell line, which is the main reason for its poor performance (61.84% in F-score). In fact, combining cell type and cell line into another single class (we call it “Newcell”) can effectively improve the performance. Our experiment shows that class “Newcell” obtain an F-score of 78.80% which is higher than the ones of cell type and cell line (75.78% and 61.84%).

Table 11 shows the performance comparison among the top JNLPBA2004, BioCreative2004 systems and ours. Through the exploitation of the contextual cues our method achieved higher performances (75.0% and 83.7% in F-score, respectively) than the systems in both NLPBA2004 and BioCreative2004 task 1A.

4. Conclusions and future work

This paper presents a CRF-based bio-entity name recognition approach and studies the methods of improving the
The JNLPBA result data are achieved from [1] and the BioCreative result data are achieved from [4,22,18], respectively.

### Table 11

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (JNLPBA)</td>
<td>77.2</td>
<td>73.0</td>
<td>75.0</td>
</tr>
<tr>
<td>Zhou (JNLPBA)</td>
<td>76.0</td>
<td>69.4</td>
<td>72.6</td>
</tr>
<tr>
<td>Finkel (JNLPBA)</td>
<td>71.6</td>
<td>68.6</td>
<td>70.1</td>
</tr>
<tr>
<td>Settles (JNLPBA)</td>
<td>70.3</td>
<td>69.3</td>
<td>69.8</td>
</tr>
<tr>
<td>Ours (BioCreative)</td>
<td>82.9</td>
<td>84.5</td>
<td>83.7</td>
</tr>
<tr>
<td>Finkel (BioCreative) [4]</td>
<td>82.8</td>
<td>83.5</td>
<td>83.2</td>
</tr>
<tr>
<td>Zhou (BioCreative) [22]</td>
<td>82.0</td>
<td>83.2</td>
<td>82.6</td>
</tr>
<tr>
<td>McDonald (BioCreative) [18]</td>
<td>86.4</td>
<td>78.7</td>
<td>82.4</td>
</tr>
</tbody>
</table>

The JNLPBA result data are achieved from [1] and the BioCreative result data are achieved from [4,22,18], respectively.

performance via the exploitation of the contextual cues. Experiment results on NLPBA2004 and BioCreative2004 task 1A datasets showed that these methods improve the recognition performance by more than 2 points in F-score.

At present machine learning techniques such as HMM, MEMM, CRFs are popular bio-entity name recognition approaches. However, their recognition performances seem to have encountered a threshold due to the irregularities and ambiguities of bio-entities nomenclature as well as the lack of large and qualified training set. In JNLPBA2004 and BioCreative2004 task 1A, using exact matching strategy, the best systems achieved an F-score of no more than 75% (72.6% and 74.3%, respectively) and it is difficult to improve the recognition performance even by one point in F-score.

On the other hand, our experiment results on JNLPBA2004 and BioCreative2004 task 1A datasets show that the exploitation of the contextual cues maybe a promising method to further improve the recognition performance. Contextual cues such as bracket pair, heuristic syntax structure and interaction words cue could provide relatively precise information about bio-entities and their class. These contextual cues as a feature of natural language abound in biomedical literature and can help improve the recognition performance on a certain degree. In addition, the exploitation of the contextual cues is not limited to any specific dataset, thus having general applicability. In fact, the annotation rules of JNLPBA2004 and BioCreative2004 task 1A datasets are not quite the same, but the method of exploitation of the contextual cues work well on both datasets.

In fact, besides the contextual cues exploited in our method, we also found other contextual cues that could be utilized. For example, in sentence “Steroid identities were confirmed by gas chromatography/mass spectrometry (GC/MS)”, “gas chromatography/mass spectrometry” and “GC/MS” are a full name–abbreviation pair. Even though the word “GC/MS” has the features of a protein (or gene) (including capital letters and “/”) but it is not likely to be a protein (or gene) since in most cases a protein (or gene) cannot “confirm” and it is more like an assay method.

To filter such false positive the related domain knowledge as well as syntax analysis tool need to be introduced. In fact, attempts to incorporate domain knowledge to improve recognition performance have been reported. McDonald and Pereira 2005 used ABGene lexicons as features and obtained an improvement of 2 points in F-score. These lexicons include general biological terms, amino acids, restriction enzymes, cell lines, organism names and non-biological terms. Zhou and Su made use of SwissProt and the alias list LocusLink as external dictionary features which improved the performance by 1.2 points in F-score.

However, most of domain knowledge ever used is only external dictionary of protein, gene and other biological entities. Our idea is to construct a domain knowledge database which includes all kind of bio-entities class, their properties and actions they can do and can be done. Then via in-domain syntactic parser, the domain knowledge database can be used to filter out false positives and recover false negatives. Interaction words cues described in our manuscript is a simple implement of our idea and it did improve the performance (0.2 and 0.5 point on JNLPBA2004 and BioCreative2004, respectively).

On the other hand, the domain knowledge database could be difficult to construct. Many online biomedical resources such as UMLS Knowledge Sources [23] (Metathesaurus and Semantic Network) and GO [24] may be utilized. Experts with backgrounds in biochemistry, genetics and molecular biology are needed to provide relative domain knowledge. In addition, at present, the performance of current in-domain syntactic parser is not good enough. All these are the problems we must solve to improve the performance by incorporating domain knowledge.

### Acknowledgments

This work is supported by grant from the Natural Science Foundation of China (No. 60373095 and 60673039) and the National High Tech Research and Development Plan of China (2006AA01Z151).

### References


