An Unsupervised Method for Extracting Domain-specific Affixes in Biological Literature

Haibin Liu   Christian Blouin   Vlado Kešelj
Faculty of Computer Science, Dalhousie University, Canada, {*haibin,cblouin,vlado*}*@cs.dal.ca*

Abstract

We propose an unsupervised method to automatically extract domain-specific prefixes and suffixes from biological corpora based on the use of PATRICIA tree. The method is evaluated by integrating the extracted affixes into an existing learning-based biological term annotation system. The system based on our method achieves comparable experimental results to the original system in locating biological terms and exact term matching annotation. However, our method improves the system efficiency by significantly reducing the feature set size. Additionally, the method achieves a better performance with a small training data set. Since the affix extraction process is unsupervised, it is assumed that the method can be generalized to extract domain-specific affixes from other domains, thus assisting in domain-specific concept recognition.

1 Introduction

Biological term annotation is a preparatory step in information retrieval in biological science. A biological term is generally defined as any technical term related to the biological domain. Considering term structure, there are two types of biological terms: single word terms and multi-word terms. Many systems (Fukuda et al., 1998; Franzn et al., 2002) have been proposed to annotate biological terms based on different methodologies in which determining term boundaries is usually the first task. It has been demonstrated (Jiampojamarn et al., 2005a), however, that accurately locating term boundaries is difficult. This is so because of the ambiguity of terms, and the peculiarity of the language used in biological literature.

(Jiampojamarn et al., 2005b) proposed an automatic biological term annotation system (ABTA) which applies supervised learning methods to annotate biological terms in the biological literature. Given unstructured texts in biological research, the annotation system first locates biological terms based on five word position classes, “Start”, “Middle”, “End”, “Single” and “Non-relevant”. Therefore, multi-word biological terms should be in a consistent sequence of classes “Start (Middle)* End” while single word terms will be indicated by the class “Single”. Word n-grams (Cavnar and Trenkle, 1994) are used to define each input sentence into classification instances. For each element in an n-gram, the system extracts feature attributes as input for creating the classification model. The extracted feature attributes include word feature patterns(e.g., Greek letters, uppercase letters, digits and other symbols), part-of-speech (POS) tag information, prefix and suffix characters. Without using other specific domain resources, the system achieves comparable results to some other state-of-the-art systems (Finkel et al., 2004; Settles, 2004) which resort to external knowledge, such as protein dictionaries. It has been demonstrated (Jiampojamarn et al., 2005b) that the part-of-speech tag information is the most effective attribute in aiding the system to annotate biological terms because most biological terms are partial noun phrases.

The ABTA system learns the affix feature by recording only the first and the last *n* characters (e.g., *n* = 3) of each word in classification instances, and the authors claimed that the *n* characters could provide enough affix information for the term annotation task. Instead of using a certain number of characters to provide affix information, however, it is more likely that a specific list of typically used prefixes and suffixes of biological words would provide more accurate information to classifying some biological terms and boundaries. We hypothesize that

BioNLP 2007: Biological, translational, and clinical language processing, pages 33–40,
Prague, June 2007. ©2007 Association for Computational Linguistics

a more flexible affix definition will improve the performance of the taks of biological term annotation.

Inspired by (Jiampojamarn et al., 2005b), we propose a method to automatically extract domain-specific prefixes and suffixes from biological corpora. We evaluate the effectiveness of the extracted affixes by integrating them into the parametrization of an existing biological term annotation system, ABTA (Jiampojamarn et al., 2005b), to evaluate the impact on performance of term annotation. The proposed method is completely unsupervised. For this reason, we suggest that our method can be generalized for extracting domain-specific affixes from many domains.

The rest of the paper is organized as follows: In section 2, we review recent research advances in biological term annotation. Section 3 describes the methodology proposed for affix extraction in detail. The experiment results are presented and evaluated in section 4. Finally, section 5 summarizes the paper and introduces future work.

2 Related Work

Biological term annotation denotes a set of procedures that are used to systematically recognize pertinent terms in biological literature, that is, to differentiate between biological terms and non-biological terms and to highlight lexical units that are related to relevant biology concepts (Nenadic and Ananiadou, 2006).

Recognizing biological entities from texts allows for text mining to capture their underlying meaning and further extraction of semantic relationships and other useful information. Because of the importance and complexity of the problem, biological term annotation has attracted intensive research and there is a large number of published work on this topic (Cohen and Hersh, 2005; Franzn et al., 2003).

Current approaches in biological term annotation can be generalized into three main categories: lexicon-based, rule-based and learning-based (Cohen and Hersh, 2005). Lexicon-based approaches use existing terminological resources, such as dictionaries or databases, in order to locate term occurrences in texts. Given the pace of biology research, however, it is not realistic to assume that a dictionary can be maintained up-to-date. A drawback of lexicon-based approaches is thus that they are not able to annotate recently coined biological terms. Rule-based approaches attempt to recover terms by developing rules that describe associated term formation patterns. However, rules are often time-consuming to develop while specific rules are difficult to adjust to other types of terms. Thus, rule-based approaches are considered to lack scalability and generalization.

Systems developed based on learning-based approaches use training data to learn features useful for biological term annotation. Compared to the other two methods, learning-based approaches are theoretically more capable to identify unseen or multi-word terms, and even terms with various writing styles by different authors. However, a main challenge for learning-based approaches is to select a set of discriminating feature attributes that can be used for accurate annotation of biological terms. The features generally fall into four classes: (1) simple deterministic features which capture use of uppercase letters and digits, and other formation patterns of words, (2) morphological features such as prefix and suffix, (3) part-of-speech features that provide word syntactic information, and (4) semantic trigger features which capture the evidence by collecting the semantic information of key words, for instances, head nouns or special verbs.

As introduced earlier, the learning-based biological term annotation system ABTA obtained an 0.705 F-score in exact term matching on Genia corpus (v3.02)[1] which contains 2,000 abstracts of biological literature. In fact, the morphological features in ABTA are learned by recording only the first and the last *n* characters of each word in classification instances. This potentially leads to inaccurate affix information for the term annotation task.

(Shen et al., 2003) explored an adaptation of a general Hidden Markov Model-based term recognizer to biological domain. They experimented with POS tags, prefix and suffix information and noun heads as features and reported an 0.661 F-score in overall term annotation on Genia corpus. 100 most frequent prefixes and suffixes are extracted as candidates, and evaluated based on difference in likelihood of part of a biological term versus not. Their method results in a modest positive improvement in recognizing biological terms. Two limitations of this method are: (1) use of only a biological corpus, so

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

that the general domain-independent affixes are not removed, and (2) a supervised process of choosing a score threshold that is used in affix selection.

(Lee et al., 2003) used prefix and suffix features coupled with a dictionary-based refinement of boundaries of the selected candidates in their experiments for term annotation. They extracted affix features in a similar way with (Shen et al., 2003). They also reported that affix features made a positive effect on improving term annotation accuracy.

In this project, we consider the quality of domain-specific affix features extracted via an unsupervised method. Successful demonstration of the quality of this extraction method implies that domain-specific affixes can be identified for arbitrary corpora without the need to manually generate training sets.

3 PATRICIA-Tree-based Affix Extraction

3.1 PATRICIA Tree

The method we propose to extract affixes from biological words is based on the use of PATRICIA tree. “PATRICIA” stands for “Practical Algorithm To Retrieve Information Coded In Alphanumeric”. It was first proposed by (Morrison, 1968) as an algorithm to provide a flexible means of storing, indexing, and retrieving information in a large file. PATRICIA tree uses path compression by grouping common sequences into nodes. This structure provides an efficient way of storing values while maintaining the lookup time for a key of O(*N*) in the worst case, where *N* is the length of the longest key. Meanwhile, PATRICIA tree has little restriction on the format of text and keys. Also it does not require rearrangement of text or index when new material is added. Because of its outstanding flexibility and efficiency, PATRICIA tree has been applied to many large information retrieval problems (Morrison, 1968).

In our project, all biological words are inserted and stored in a PATRICIA tree, using which we can efficiently look up specific biological word or extract biological words that share specified affixes and calculated required statistics.

3.2 Experiment Design

In this work, we have designed the experiments to extract domain-specific prefixes and suffixes of biological words from a biological corpus, and investigate whether the extracted affix information could facilitate better biological term annotation. The overall design of our experiments consists of three major processes: affix extraction, affix refining and evaluation of experimental results. It is seen that every node in PATRICIA tree contains exactly one string of 1 or more characters, which is the preceding substring of its descendant nodes. Meanwhile, every word is a path of substrings from the root node to a leaf. Therefore, we propose that every substring that can be formed from traversing the internal nodes of the tree is a potential affix.

In the affix extraction process, we first populate a PATRICIA tree using all words in the combined corpus*(CC)* of a Biological Corpus *(BC)* and a General English Corpus *(GEC). GEC* is used against *BC* in order to extract more accurate biological affix information. Two PATRICIA trees are populated separately for extracting prefixes and suffixes. The suffix tree is based on strings derived by reversing all the input words from the combined corpus. All the potential prefixes and suffixes are then extracted from the populated PATRICIA trees.

In the affix refining process, for each extracted potential affix, we compute its joint probability of being both an English affix and a biological affix, $P(D = \textit{Biology}, A = \textit{Yes}|\textit{PA})$, where *D* stands for *Domain*, *A* stands for *Affix* and *PA* represents *Potential Affix*. This joint probability can be further decomposed as shown in Eq.(1). In the formula, $P\left(A = \textit{Yes}|PA\right)$ denotes the probability that a given potential affix is a true English affix while $P(D = \textit{Biology}|A = \textit{Yes}, \textit{PA})$ refers to the probability that a given English affix is actually a biological affix.

$$P(D = \textit{Biology}, A = \textit{Yes}|PA) = \\ P(D{=}\textit{Biology}|A{=}\textit{Yes}, \textit{PA}) \times P(A{=}\textit{Yes}|\textit{PA}) \quad (1)$$

To calculate $P\left(A = \textit{Yes}|PA\right)$, the probabilities of prefixes and suffixes are measured separately. In linguistics, a prefix is described as a type of affix that precedes the morphemes to which it can attach (Soanes and Stevenson, 2004). Simply speaking, a prefix is a substring that can be found at the beginning of a word. Our functional definition of a prefix is a substring which precedes words existing in the English language. This can be done by enumerating, for each node, all descendant substring and assessing their existence as stand-alone words. For example, “radioimmunoassay”, “radioiodine” and “radio-labeled" are three words and have a common starting string "radio". If we take out the remaining part of each word, three new strings are obtained, "immunoassay", "iodine" and "labeled". Since all the input words are already stored in PATRICIA tree, we lookup these three strings in PATRICIA tree and find that "immunoassay", "iodine" and "labeled" are also meaningful words in the tree. This indicates that "radio" is a prefix among the input words. On the other hand, it is obvious that "radioimmunoassay" and "radioiodine" share another string "radioi". However, "mmunoassay" and "odine" are not meaningful words due to their absence in the PATRICIA tree. This suggests that "radioi" is not a prefix.

For each extracted potential prefix, $P\left(A = \mathit{Yes}|PA\right)$ is computed as the proportion of strings formed by traversing all descendant nodes that are meaningful terms. In our experiments, the measure of determining a string meaningful is to look up whether the string is an existing word present in the built prefix PATRICIA tree. Algorithm 1 shows the procedure of populating a PATRICIA tree and calculating $P\left(A = \mathit{Yes}|PA\right)$ for each potential prefix.

Algorithm 1 $P\left(A = \mathit{Yes}|PA\right)$ for Prefix

```
Input: words $(w) \in$ Combined Corpus $(CC)$
Output: $P\left(A = \mathit{Yes}|PA\right)$ for each potential prefix
  $PT = \emptyset$                         //$PT$ : Patricia Trie
  for all words $w \in CC$ do
      $PT \leftarrow$ Insert$(w)$        //Populating Patricia Trie
  for all nodes $n_i \in PT$ do
      $PA \leftarrow$ String$(n_i)$           //Concatenate strings
                                   // in nodes from root to $n_i$,
                                   // which is a potential prefix
      $T_{PA} \leftarrow$ PrefixSearch$(PA)$
      //$T_{PA}$ : all words $w \in CC$ beginning with $PA$
      $score \leftarrow 0$
      for all words $w \in T_{PA}$ do
          if Extrstr$(PA, w)$ in $PT$ then
              //Extrstr() returns the remaining string
              // of $w$ without $PA$
              $score$ ++
      $P(A = \mathit{Yes}|PA) \leftarrow score/|T_{PA}|$
      //$|T_{PA}|$ is the number of words in $T_{PA}$
```

Likewise, in linguistics a suffix is an affix that follows the morphemes to which it can attach (Soanes and Stevenson, 2004). Simply speaking, a suffix of a word is a substring exactly matching the last part of the word. Similar to the idea of calculating $P\left(A = \mathit{Yes}|PA\right)$ for potential prefix, we conjecture that the extracted potential suffix could be a reasonable English suffix if the inverted strings formed from traversing the descendant nodes of the potential suffix in the suffix PATRICIA tree are meaningful words. For instance, "Calcium-dependent", "Erythropoietin-dependent" and "Ligand-dependent" share a common ending string "-dependent". Since the remaining strings of each word, "Calcium", "Erythropoietin" and "Ligand" can be found in the "forward" PATRICIA tree, "-dependent" is a potentially useful suffix.

However, it is often observable that some English words do not begin with another meaningful word but a typical prefix, for example, "alpha-bound" and "pro-glutathione". It is known that "-bound"and "-glutathione" are good suffixes in biology. "alpha" and "pro", however, are not meaningful words but typical prefixes, and in fact have been extracted when calculating $P\left(A = \mathit{Yes}|PA\right)$ for potential prefix. Therefore, in order to detect and capture such potential suffixes, we further assume that if a word begins with a recognized prefix instead of another meaningful word, the remaining part of the word still has the potential to be an informative suffix. Therefore, strings "-bound" and "-glutathione" can be successfully extracted as potential suffixes. In our experiments, an extracted potential prefix is considered a recognized prefix if its $P\left(A = \mathit{Yes}|PA\right)$ is greater than 0.5.

To calculate $P(D = \mathit{Biology}|A = \mathit{Yes}, PA)$, it is necessary to first determine true English affixes from extracted potential affixes. In our experiments, we consider that an extracted potential prefix or suffix is a recognized affix only if its $P\left(A = \mathit{Yes}|PA\right)$ is greater than 0.5. It is also necessary to consider the biological corpus *BC* and the general English corpus *GEC* separately. It is assumed that a biology related affix tends to occur more frequently in words of *BC* than *GEC*. Eq.(2) is used to estimate $P\left(D = \mathit{Biology}|A = \mathit{Yes}, PA\right)$.

$$P\left(D = \mathit{Biology}|A = \mathit{Yes}, PA\right) = \\ \left(\#Words\ with\ PA\ in\ BC/Size\left(BC\right)\right)/ \\ \left(\#Words\ with\ PA\ in\ BC/Size\left(BC\right) + \right. \\ \left. \#Words\ with\ PA\ in\ GEC/Size\left(GEC\right)\right), \quad (2)$$

where only *PA* with $P\left(A = \textit{Yes}|PA\right)$ greater than 0.5 are used, and the number of words with a certain *PA* is further normalized by the size of each corpus.

Finally, the joint probability of each potential affix, $P\left(D = \textit{Biology}, A = \textit{Yes}|PA\right)$, can be used to parametrize a word beginning or ending with *PA*.

In the evaluation process of our experiments, the prefix-suffix pair with maximum joint probability values is used to parametrize a word. Therefore, each word in *BC* has exactly two values as affix feature: a joint probability value for its potential prefix and a joint probability value for its potential suffix. We then replace the original affix feature of ABTA system with our obtained joint probability values, and investigate whether these new affix information leads to equivalent or better term annotation on *BC*.

4 Results and Evaluation

4.1 Dataset and Environment

For our experiments, it is necessary to use a corpus that includes widely used biological terms and common English words. This dataset, therefore, will allow us to accurately extract the information of biology related affixes. As a proof-of-concept prototype, our experiments are conducted on two widely used corpora: Genia corpus (v3.02) and Brown corpus[2].The Genia version 3.02 corpus is used as the biological corpus *BC* in our experiments. It contains 2,000 biological research paper abstracts. They were selected from the search results in the MEDLINE database[3], and each biological term has been annotated into different terminal classes based on the opinions of experts in biology. Used as the general English corpus *GEC*, Brown corpus includes 500 samples of common English words, totalling about a million words drawn from 15 different text categories.

All the experiments were executed on a Sun Solaris server Sun-Fire-880. Our experiments were mainly implemented using Perl and Python.

4.2 Experimental Results

We extracted 15,718 potential prefixes and 21,282 potential suffixes from the combined corpus of Genia and Brown. Among them, there are 2,306 potential prefixes and 1,913 potential suffixes with joint probability value $P\left(D = \textit{Biology}, A = \textit{Yes}|PA\right)$ greater than 0.5. Table 1 shows a few examples of extracted potential affixes whose joint probability value is equal to 1.0. It is seen that most of these potential affixes are understandable biological affixes which directly carry specific semantic meanings about certain biological terms. However, some substrings are also captured as potential affixes although they may not be recognized as "affixes" in linguistics, for example "adenomyo" in prefixes, and "mopoiesis" in suffixes. In Genia corpus, "adenomyo" is the common beginning substring of biological terms "adenomyoma", "adenomyosis" and "adenomyotic" , while "plasias" is the common ending substring of biological terms "neoplasias" and "hyperplasias". The whole list of extracted potential affixes is available upon request.

In order to investigate whether the extracted affixes improves the performance of biological term annotation, it is necessary to obtain the experimental results of both original ABTA system and the ABTA system using our extracted affix information. In ABTA, the extraction of feature attributes is performed on the whole 2000 abstracts of Genia corpus, and then 1800 abstracts are used as training set while the rest 200 abstracts are used as testing set. The evaluation measures are precision, recall and F-score. C4.5 decision tree classifier (Alpaydin, 2004) is reported as the most efficient classifier which leads to the best performance among all the classifiers experimented in (Jiampojamarn et al., 2005b). Therefore, C4.5 is used as the main classifier in our experiments. The experimental results of ABTA system with 10 fold cross-validation based on different combinations of the original features are presented in Table 2 in which feature *"WFP"* is short for Word Feature Patterns, feature *"AC"* denotes Affix Characters, and feature *"POS"* refers to POS tag information. The setting of parameters in the experiments with ABTA is: the word n-gram size is 3, the number of word feature patterns is 3, and the number of affix characters is 4. We have reported the F-score and the classification accuracy of the experiments in the table. It is seen that there is a tendency with the experimental performance that for a multi-word biological term, the middle position is most difficult to detect while the ending position is generally easier to be identified than the starting position. The assumed reason for this tendency is that for multi-

[2] http://clwww.essex.ac.uk/w3c/corpus_ling/

[3] http://www.ncbi.nlm.nih.gov/PubMed/

Potential Prefixes				Potential Suffixes			
13-acetate	adenomyo	3-kinase	platelet	-T-cell	-alpha-activated	cytoid	-methyl
B-cell	Rel/NF-kappaB	CD28	pharmaco	-coated	mopoiesis	-bearing	lyse
endotoxin	anti-CD28	HSV-1	adenovirus	-expressed	-nonresponsive	-kappaB-mediated	-receptor
I-kappaB	VitD3	ligand	chromatin	-inducer	coagulant	-globin-encoding	glycemia
macrophage	cytokine	N-alpha-tosyl-L	hemoglobin	plasias	-soluble	-immortalized	racrine

Table 1: Examples of Extracted Potential Affixes with Joint Probability Value 1.0

word biological terms, many middle words of are seemingly unrelated to biology domain while many ending words directly indicate their identity, for instances, “receptor”, “virus” or “expression”.

Table 3 shows the experimental results of ABTA system after replacing the original affix feature with our obtained joint probability values for each word in Genia corpus. *“JPV”* is used to denote Joint Probability Values. It is seen that based on all three features the system achieves a classification accuracy of 87.5%, which is comparable to the results of the original ABTA system. However, the size of the feature set of the system is significantly reduced, and the classification accuracy of 87.5% is achieved based on only 18 parameters, which is 1/2 of the size of the original feature set. Meanwhle, the execution time of the experiments generally reduces to nearly half of the original ABTA system (e.g., reduces from 4 hours to 1.7 hours). Furthermore, when the feature set contains only our extracted affix information, the system reaches a classification accuracy of 81.46% based on only 6 parameters. It is comparable with the classification accuracy achieved by using only POS information in the system. In addition, Table 3 also presents the experimental results when our extracted affix information is used as an addtional feature to the original feature set. It is expected that the system performance is further improved when the four features are applied together. However, the size of the feature set increases to 42 parameters, which increases the data redundancy. This proves that the extracted affix information has a positive impact on locating biological terms, and it could be a good replacement of the original affix feature.

Moreover, we also evaluated the performance of the exact matching biological term annotation based on the obtained experimental results of ABTA system. The exact matching annotation in ABTA system is to accurately identify every biological term, including both multi-word terms and single word terms, therefore, all the word position classes of a term have to be classified correctly at the same time. An error occurring in any one of “Start” “Middle” and “End” classes leads the system to annotate multi-word terms incorrectly. Consequently, the accumulated errors will influence the exact matching annotation performance. Table 4 presents the exact matching annotation results of different combination of features based on 10 fold cross-validation over Genia corpus. It is seen that after replacing the original affix feature of ABTA system with our obtained joint probability values for each word in Genia corpus, the system achieves an 0.664 F-score on exact matching of biological term annotation, comparable to the exact matching performance of the original ABTA system. In addition, when the feature set contains only our extracted affix information, the system reaches an 0.536 F-score on exact matching. Although it is a little lower than the exact matching performance achieved by using only the original affix features in the system, the feature set size of the system is significantly reduced from 24 to 6.

In order to further compare our method with the original ABTA system, we attempted eleven different sizes of training data set to run the experiments separately based on our method and the original ABTA system. They can then be evaluated in terms of their performance on each training set size. These eleven different training set sizes are: 0.25%, 0.5%, 1%, 2.5%, 5%, 7.5%, 10%, 25%, 50%, 75% and 90%. For instance, 0.25% denotes that the training data set is 0.25% of Genia corpus while the rest 99.75% becomes the testing data set for experiments. It is observed that there are about 21 paper abstracts in training set when its size is 1% , and 52 abstracts when its size is 2.5%. It is expected that larger training set size leads to better classification accuracy of experiments.

For each training set size, we randomly extracted 10 different training sets from Genia corpus to run the experiments. We then computed the *mean classification accuracy (MCA)* of 10 obtained classification accuracies. Figure 1 was drawn to illustrate the distribution of MCA of each training set size

Feature	F-Measure					Classification	#
sets	Start	Middle	End	Single	Non	Accuracy (%)	Parameters
WFP	0.467	0.279	0.495	0.491	0.864	74.59	9
AC	0.709	0.663	0.758	0.719	0.932	85.67	24
POS	0.69	0.702	0.775	0.67	0.908	83.96	3
WFP+AC	0.717	0.674	0.762	0.730	0.933	86.02	33
WFP+POS	0.726	0.721	0.793	0.716	0.923	85.96	12
AC+POS	0.755	0.741	0.809	0.732	0.930	87.14	27
WFP+AC+POS	0.764	0.745	0.811	0.749	0.933	**87.59**	**36**

Table 2: Experimental Results of Original ABTA System

Feature	F-Measure					Classification	#
sets	Start	Middle	End	Single	Non	Accuracy (%)	Parameters
JPV	0.652	0.605	0.713	0.602	0.898	**81.46**	**6**
WFP+JPV	0.708	0.680	0.756	0.699	0.919	84.84	15
JPV+POS	0.753	0.740	0.805	0.722	0.928	86.92	9
WFP+JPV+POS	0.758	0.749	0.809	0.74	0.933	**87.50**	**18**
WFP+AC+POS+JPV	0.767	0.746	0.816	0.751	0.934	87.77	42

Table 3: Experimental Results of ABTA System with Extracted Affix Information

for both methods, with the incremental proportion of training data. It is noted in Figure 1 that the change patterns of MCA obtained by our method and the original ABTA system are similar. It is also seen that our method achieves marginally better classification performance when the proportion of training dat

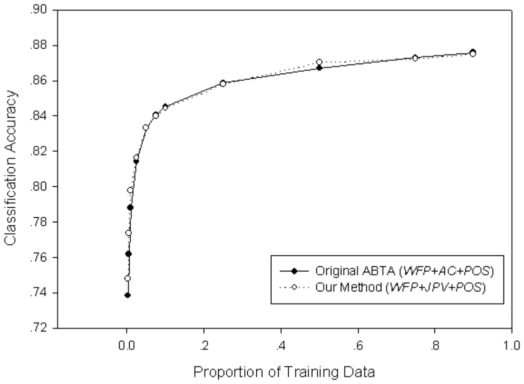

Figure 1: MCA Distribution

In order to determine if the classification performance difference between our method and the original ABTA system is statistically significant, we performed one-tailed t-Test (Alpaydin, 2004) on the classification results with our hypothesis that MCA of our proposed method is higher than MCA of original ABTA system. The significance level α is set to be the conventional value 0.05. As a result, the classification performance difference between two methods is statistically significant when the proportion of training data is 0.25%, 0.5%, 1% or 2.5%. Table 5 shows the *P* values of t-Test results for the various training set sizes. This demonstrates that the ABTA system adopting our method outperforms the original ABTA system in classification accuracy when the proportion of training data is lower than 2.5% of Genia corpus, and achieves comparable classification performance with the original ABTA system when the proportion continuously increases.

One-tailed	Training set size			
t-Test	0.25%	0.5%	1%	2.5%
P value	0.0298	0.0006	0.0002	0.0229

Table 5: One-tailed t-Test Results

5 Conclusions

In this paper, we have presented an unsupervised method to extract domain-specific prefixes and suffixes from the biological corpus based on the use of PATRICIA tree. The ABTA system (Jiampojamarn et al., 2005b) adopting our method achieves an overall classification accuracy of 87.5% in locating biological terms, and derives an 0.664 F-score in exact term matching annotation, which are all comparable to the experimental results obtained by the original ABTA system. However, our method helps the system significantly reduce the size of feature set and thus improves the system efficiency. The system also obtains a classification accuracy of 81.46% based only on our extracted affix information. This

Feature	Exact Matching Annotation			#
sets	Precision	Recall	F-score	Parameters
AC	0.548	0.571	0.559	24
WFP+AC+POS	0.661	0.673	0.667	36
JPV	0.527	0.545	0.536	6
WFP+JPV+POS	0.658	0.669	0.664	18

Table 4: Exact Matching Annotation Performance

demonstates that the affix information acheived by the proposed method is important to accurately locating biological terms.

We further explored the reliability of our method by gradually increasing the proportion of training data from 0.25% to 90% of Genia corpus. One-tailed t-Test results confirm that the ABTA system adopting our method achieves more reliable performance than the original ABTA system when the training corpus is small. The main result of this work is thus that affix features can be parametrized from small corpora at no cost in performance.

There are some aspects in which the proposed method can be improved in our future work. We are interested in investigating whether there exists a certain threshold value for the joint probability which might improve the classification accuracy of ABTA system to some extent. However, this could import supervised elements into our method. Moreover, we would like to incorporate our method into other published learning-based biological term annotation systems to see if better system performance will be achieved. However, superior parametrization will improve the annotation performance only if the affix information is not redundant with other features such as POS.

References

Ethem Alpaydin. 2004. *Introduction to Machine Learning*. MIT Press.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proc. SDAIR-94, 3rd Ann. Symposium on Doc. Analysis and Inf. Retr.*, pages 161–175, Las Vegas, USA.

Aaron Michael Cohen and William R. Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, 5(1):57–71.

Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Gail Sinclair, and Christopher Manning. 2004. Exploiting context for biomedical entity recognition: From syntax to the web. In *Joint wsh. on NLP in Biomedicine and its Applications (JNLPBA-2004)*.

Kristofer Franzn, Gunnar Eriksson, Fredrik Olsson, Lars Asker Per Lidn, and Joakim Cster. 2002. Protein names and how to find them. *International Journal of Medical Informatics special issue on NLP in Biomedical Applications*, pages 49–61.

Kristofer Franzn, Gunnar Eriksson, Fredrik Olsson, Lars Asker Per Lidn, and Joakim Cster. 2003. Mining the Biomedical Literature in the Genomic Era: An Overview. *J. Comp. Biol.*, 10(6):821–855.

K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. Toward information extraction: Identifying protein names from biological papers. In *the Pacific Symposium on Biocomputing*, pages 707–718.

Sittichai Jiampojamarn, Nick Cercone, and Vlado Kešelj. 2005a. Automatic Biological Term Annotation Using N-gram and Classification Models. Master's thesis, Faculty of Comp.Sci., Dalhousie University.

Sittichai Jiampojamarn, Nick Cercone, and Vlado Kešelj. 2005b. Biological Named Entity Recognition using N-grams and Classification Methods. In *Conf. of the Pacific Assoc. for Computational Linguistics, PACLING'05*, Tokyo, Japan.

Ki-Joong Lee, Young-Sook Hwang, and Hae-Chang Rim. 2003. Two-phase biomedical NE recognition based on SVMs. In *Proc. of the ACL 2003 workshop on Natural language processing in biomedicine*, pages 33–40, Morristown, NJ, USA. ACL.

Donald R. Morrison. 1968. Patricia - Practical Algorithm To Retrieve Information Coded in Alphanumeric. *Journal of the ACM*, 15(4):514–534.

Goran Nenadic and Sophia Ananiadou. 2006. Mining semantically related terms from biomedical literature. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(1):22 – 43.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and novel feature sets. In *Joint wsh. on NLP in Biomedicine and its Applications (JNLPBA-2004)*.

Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2003. Effective adaptation of a Hidden Markov Model-based named entity recognizer for biomedical domain. In *Proc. of the ACL 2003 wsh. on NLP in Biomedicine*, pages 49–56, Morristown, NJ, USA.

Catherine Soanes and Angus Stevenson. 2004. *Oxford Dictionary of English*. Oxford University Press.