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

Haibin Liu
Christian Blouin
Vlado Keselj

Technical Report CS-2008-01

January 16, 2008

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

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

Haibin Liu[1], Christian Blouin[1] and Vlado Kešelj[1]

[1]Faculty of Computer Science, Dalhousie University

Email: H. Liu— haibin@cs.dal.ca; C. Blouin — cblouin@cs.dal.ca; V. Kešelj — vlado@cs.dal.ca;

Abstract

Background: 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.

Results: 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.

Conclusions: 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.

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. Annotating important terms in unstructured biological text such as molecules, genes and proteins enables the discovery of patterns and semantic relations in the text. Many systems [1–3] 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 [4], 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. Consequently, different terms that are possibly semantically unrelated could be generated due to erroneously located boundaries.

Jiampojamarn *et al.* [5] 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". The system then associates each located term with biological concept markers indicating whether the term is a "protein", "DNA" or "RNA" instance name. Word n-grams [6] 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 [3, 7, 8] which resort to external knowledge, such as protein dictionaries. It has been demonstrated [5] 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 and the POS tag information reduces the confusion over other tags such as determiner, verb, and preposition. However, in ABTA, the affix characters extracted from words did not improve the overall system performance as significantly as the part-of-speech tag information.

In fact, 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. However, it is often observed that most biological terms tend to employ longer affixes which carry specific semantic meanings about the terms. Instead of using a certain number of characters to provide affix information, 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 therefore hypothesize that a more flexible affix definition will improve the performance of the task of biological term annotation.

Inspired by the work of Jiampojamarn *et al.* [5], 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 [5], to evaluate the impact on performance of term annotation. The proposed method is completely unsupervised. For this reason, we assume that our method can be generalized for extracting domain-specific affixes from many domains. Without prior knowledge about a certain domain, the extracted affixes could help to recognize specific domain terms, understand domain concepts, and further facilitate semantic processing of the text in the domain.

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.

Related Work

A term usually corresponds to an author's textual representation of a particular concept. It is not easy to understand an article without precise identification of terms that are used to communicate knowledge. 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 [9].

Recognizing biological entities from texts allows for text mining to capture their underlying meaning and further extraction of semantic relationships and other useful information. Automating this process enables the large-scale processing of biomedical literature. 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 [10, 11].

Current approaches in biological term annotation can be generalized into three main categories: lexicon-based, rule-based and learning-based [10]. 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 reliably 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. The classification technique allows systems to automatically learn from examples, and then produce solutions to similar problems using the learned model. 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 term instances. 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 extracts feature attributes including word feature patterns, part-of-speech (POS) tag information, prefix and suffix characters. Based on these features, the system creates the classification model and uses the model to annotate terms in biological literature. The system obtained an 0.705 F-score in exact term matching on Genia corpus (v3.02) [12] 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.* [13] 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 [12]. In their experiments, 100 most frequent prefixes and suffixes were extracted as candidates, and then evaluated in terms of a score obtained by a proposed formula which assumes that a particular prefix or suffix that is most likely inside biological terms would be least likely outside biological terms. A threshold was chosen and considered as a good evidence for distinguishing biological terms. They reported that prefix and suffix information showed a modest positive effect on recognizing biological terms. The proposed formula, however, distinguishes between biological and non-biological affixes based solely on a biological corpus. This could potentially bias the results since not all frequent affixes occurring in biological terms of Genia corpus are good molecular biology affixes. Meanwhile, in order to choose a proper threshold, it is necessary to scrutinize all the extracted prefixes and suffixes. This makes the process of extracting affix features supervised.

Lee *et al.* [14] 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.* [13]. 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.

Methodology: PATRICIA-Tree-based Affix Extraction

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 [15] 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 [15].

Figure 1 illustrates a simple example of the growth of a PATRICIA tree under a sequence of insertions. Suppose "ababb", "ababa", "ba" and "aaabba" are four words to be inserted into PATRICIA tree. Initially, the tree is empty. The word "ababb" is first inserted as a node of the tree. In order to insert the next word "ababa", it is necessary to compare it with the existing node containing the word "ababb". The only difference of the two words is the fifth character. Thus, the common substring "abab" is split out from both words to form a new node, while the remaining characters of each word, "a" and "b", become two descendant nodes of this node. Similarly, it is necessary to compare the next inserted word "ba" with "ababb". As they are different from the first character, a new node for "ba" is then created from the root of the tree. Finally, since the common character of the last inserted word "aaabba" and "ababb" is "a", it is then split out as a new node.

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 calculate required statistics.

Terminology

In order to clearly describe our proposed method, it is necessary to introduce and clearly define some terminology. In particular, we use several notions of suffixes and prefixes, commonly called affixes, and we define them using basic formal language terminology.

We assume that text is composed of characters from a finite alphabet. The **alphabet** is a finite set of symbols Σ. In practice, elements of alphabet are ASCII characters. A **string** is a finite sequence of characters from Σ, which can be empty. The **length of a string** w is the number of characters in the sequence w, and it is denoted $|w|$. A **text,** which is in our case a scientific article, is simply a long string. Two strings s and t can be **concatenated** by appending one string to another, which is denoted as st or $s \cdot t$. A s is a **substring** of string t if $t = xsy$ for some strings x and y. It is assumed that a string can be an empty string.

The standard definition of *prefix* and *suffix* in the formal languages theory is that a string s is a prefix of string t if $t = sx$ for some string x, and s is a suffix of string t if $t = xs$ for some

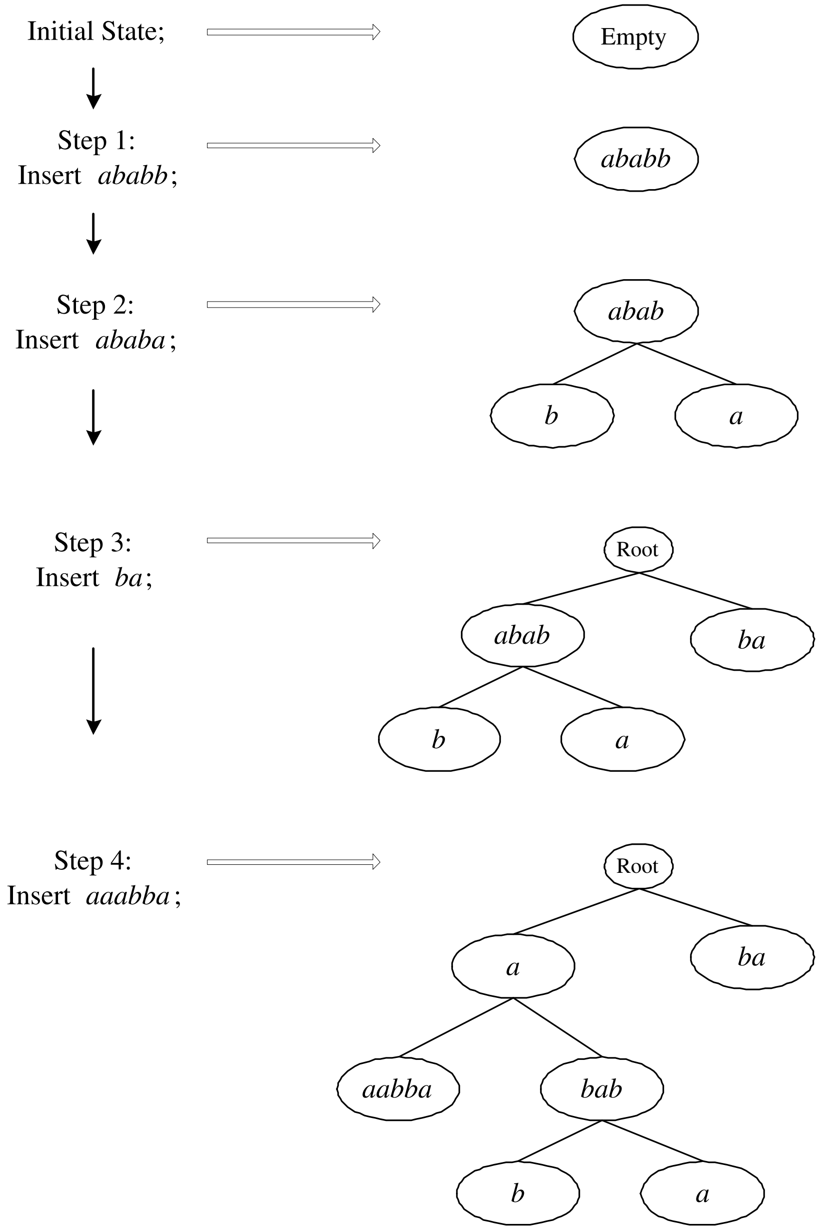

Figure 1: Growth of a PATRICIA Tree

string x. However, since we are going to define terms *prefix* and *suffix* in a different way in our proposed method, we will call these standard notions of *prefix* and *suffix*, **general prefix** and **general suffix,** or if we want to include both terms under one term, **general affix.**

We define **words** as the maximal substrings of a text consisting only of the characters from a distinguished subset of Σ, which is the set of all letters, digits, and some special signs such as the hyphen sign '-' and the slash sign '/'. A **corpus** is a finite set of texts. We use two corpora: the Brown corpus as a general English corpus, and the Genia corpus as a domain-specific biological corpus. We say that a word w occurs in a corpus $\mathcal{C}$ if there is a text $T \in \mathcal{C}$ such that w occurs in T.

The goal of our method is to extract a set of domain-specific affixes (i.e., prefixes and suffixes), so that we can use them as a special feature in biological term detection. Being a **domain-specific affix** is not a deterministic feature of a string; instead, we use a probabilistic model to predict a probability that given an arbitrary occurrence of an affix in a text, it has a domain-specific function, i.e., its occurrence is related to biological domain in some way. The task becomes more difficult by performing this function in a very unsupervised way in the sense that we do not use information about annotated biological terms, but treat the domain-specific corpus as unmarked, uniform text.

The first step is to identify potential affixes, which we define in the following way:

Definition 1 (Potential Affix) *Given a corpus $\mathcal{C}$ over an alphabet Σ, a string p, $|p| \geq 2$, is a* **potential prefix** *if there are two distinct words w_1 and w_2 occurring in $\mathcal{C}$, such that $w_1 = px_1$ and $w_2 = px_2$ for two strings x_1 and x_2. Similarly, a string s, $|s| \geq 2$, is a* **potential suffix** *if there are two distinct words w_1 and w_2 occurring in $\mathcal{C}$, such that $w_1 = x_1s$ and $w_2 = x_2s$ for two strings x_1 and x_2. A* **potential affix** *is a potential prefix or potential suffix.*

For each node n in a PATRICIA tree we can construct a string p_n by concatenating all strings associated with nodes in the path from root to n. If we store all words from a corpus in a PATRICIA tree, then the strings p_n for all nodes n are exactly all potential prefixes of the corpus. This is an efficient algorithm for finding all potential prefixes of a corpus. The potential suffixes are extracted in a similar way by reversing all words before adding them to the tree, and then by reversing the extracted potential prefixes at the end.

In the second step affixes are extracted from potential affixes. The notions of prefix, suffix, and affix are defined as follows:

Definition 2 (Affix) *A string p is a* **prefix** *in a corpus $\mathcal{C}$ if it is a potential prefix, and there are two words w_1 and w_2 occurring in the corpus $\mathcal{C}$ such that $w_1 = pw_2$. A string s is a* **suffix** *in corpus $\mathcal{C}$ if it is a potential suffix and there is a word w_1, and a word w_2 or prefix w_2 in the corpus $\mathcal{C}$ such that $w_1 = w_2s$. An* **affix** *is a prefix or suffix.*

For example, "radio" is a potential prefix due to existence of the three words: "radioimmunoassay", "radioiodine" and "radiolabeled". Since "immunoassay" is also a word, "radio" is a prefix. Actually, “iodine” and “labeled” are words as well. On the other hand, “radioi” is a potential prefix shared by the words “radioimmunoassay” and “radioiodine”, but it is not a prefix since neither “mmunoassay” nor “odine” is a word. Similarly, “Calcium-dependent”, “Erythropoietin-dependent” and “Ligand-dependent” share a potential suffix “-dependent”, which is also a suffix since “Calcium”, “Erythropoietin” and “Ligand” are words. Additionally, the word “pro-glutathione” has the suffix “-glutathione” not because “pro” is a word, but because “pro” is a prefix. In this sense suffixes are not handled in a completely equivalent way as prefixes.

Probabilistic Model

In order to determine the property of an affix to be a *domain-specific affix*, we use a probabilistic model. We in fact use two probabilistic models—one for prefixes and one for suffixes. Since they are analogue methods, we will use only the prefix model in our explanation.

A random event in our model is occurrence of a general prefix in a specific instance of a word in a corpus, i.e., it is an instance of a word $w = px$, where p is a general prefix. By an “instance of a word w” we mean an occurrence of a word w at a specific position in a specific text of the corpus; otherwise, the same word may occur at many positions and in many texts. We assume a uniform distribution of these events.

We define the following random variables for this event:

w is the word occurring in the event.

p is the general prefix occurring in the event.

$\mathit{PA} \in \{\top, \bot\}$ is the property of the general prefix p to be a potential prefix or not.

$A \in \{\top, \bot\}$ is the property of the general prefix p to occur as a prefix in the word w, or not.

$D \in \{\mathit{Biology}, \mathit{Non\text{-}B}\}$ is the property that occurrence of the general prefix p in this context is specific for biological domain, or not. In many instances this property may require a human domain expert to be determined, and even then it may not be clear, but we assume that it exists.

Among all random events, we are interested only in those in which p is a potential prefix. The goal of our model is to find the expected probability that a given p will be an English affix and actually a biological domain-specific affix; i.e., we want to find the joint probability:

$$P(D = \mathit{Biology}, A = \top \mid \mathit{PA} = \top, p)$$

Using the chain-rule property, this probability can be expressed as:

$$\begin{aligned} & P(D = \mathit{Biology}, A = \top \mid \mathit{PA} = \top, p) = \\ & \quad = \; P(D = \mathit{Biology} \mid A = \top, \mathit{PA} = \top, p) \cdot P(A = \top \mid \mathit{PA} = \top, p) \end{aligned} \tag{1}$$

$P(A = \top \mid \textit{PA} = \top, p)$ denotes the probability that a given potential prefix is a true English prefix while $P(D = \textit{Biology} \mid A = \top, \textit{PA} = \top, p)$ refers to the probability that a given English prefix is actually a biology related affix.

The value of $P(A = \top \mid \textit{PA} = \top, p)$ is estimated using the maximum likelihood estimation (MLE) based on the combined corpus*(CC)* of a Biological Corpus *(BC)* and a General English Corpus *(GEC)*; i.e., it is the number of prefix occurrences of p divided by the number of times p occurs as potential prefix in the corpus. *GEC* is used against *BC* in order to extract more accurate biological affixes. Algorithm 1 shows the procedure of populating a PATRICIA tree and calculating the value of $P(A = \top \mid \textit{PA} = \top, p)$ for each potential prefix.

```
Algorithm 1  $P(A = \top \mid \textit{PA} = \top, p)$ for Prefix
Input:  words $(w) \in$ Combined Corpus $(CC)$
Output:  $P(A = \top \mid \textit{PA} = \top, p)$ for each potential prefix
  PT = $\emptyset$                                         //$PT$ : Patricia tree
  for all words $w \in CC$ do
      PT ← Insert($w$)                                     //Populating Patricia tree
  for all internal nodes $n_i \in$ PT do
      PA ← String($n_i$)        //Concatenate strings in nodes from root to $n_i$, which is a potential prefix
      $T_{PA}$ ← PrefixSearch(PA)
      //$T_{PA}$ : all words $w \in CC$ beginning with PA
      score ← 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
              $score$ ++
      $P(A = \top \mid \textit{PA} = \top, p)$ ← score/$|T_{PA}|$
      //$|T_{PA}|$ is the number of words in $T_{PA}$
```

In order to measure the value of $P(D = \textit{Biology} \mid A = \top, \textit{PA} = \top, p)$, it is necessary to first determine true English prefixes from potential prefixes. A potential prefix is considered a prefix in our method if $P(A = \top \mid \textit{PA} = \top, p) \geq 0.5$. It is also necessary to consider the biological corpus *BC* and the general English corpus *GEC* separately. It is assumed that a biology related prefix tends to occur more frequently in words of *BC* than *GEC*. Eq.(2) is used to estimate the value of $P(D = \textit{Biology} \mid A = \top, \textit{PA} = \top, p)$.

$$\begin{aligned} &P(D = \textit{Biology} \mid A = \top, \textit{PA} = \top, p) = \\ &\quad (\#Words\ with\ \textit{PA}\ in\ BC/Size\,(BC))/ \\ &\quad (\#Words\ with\ \textit{PA}\ in\ BC/Size\,(BC) + \\ &\quad \#Words\ with\ \textit{PA}\ in\ GEC/Size\,(GEC)), \end{aligned} \tag{2}$$

where only *PA* with $P(A = \top \mid \textit{PA} = \top, p) \geq 0.5$ are used, and the number of words with a certain prefix is further normalized by the size of each corpus.

These probability values can be efficiently calculated using PATRICIA tree. Based on these values and the formula given above, we can calculate the joint probability $P(D = \textit{Biology}, A = \top \mid \textit{PA} = \top, p)$ for each potential prefix p. The analogue joint probability is calculated for all potential suffixes.

Experiment Design

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. Figure 2 illustrates the overall design of our experiments, which consists of three major processes: affix extraction, affix refining, and evaluation of experimental results.

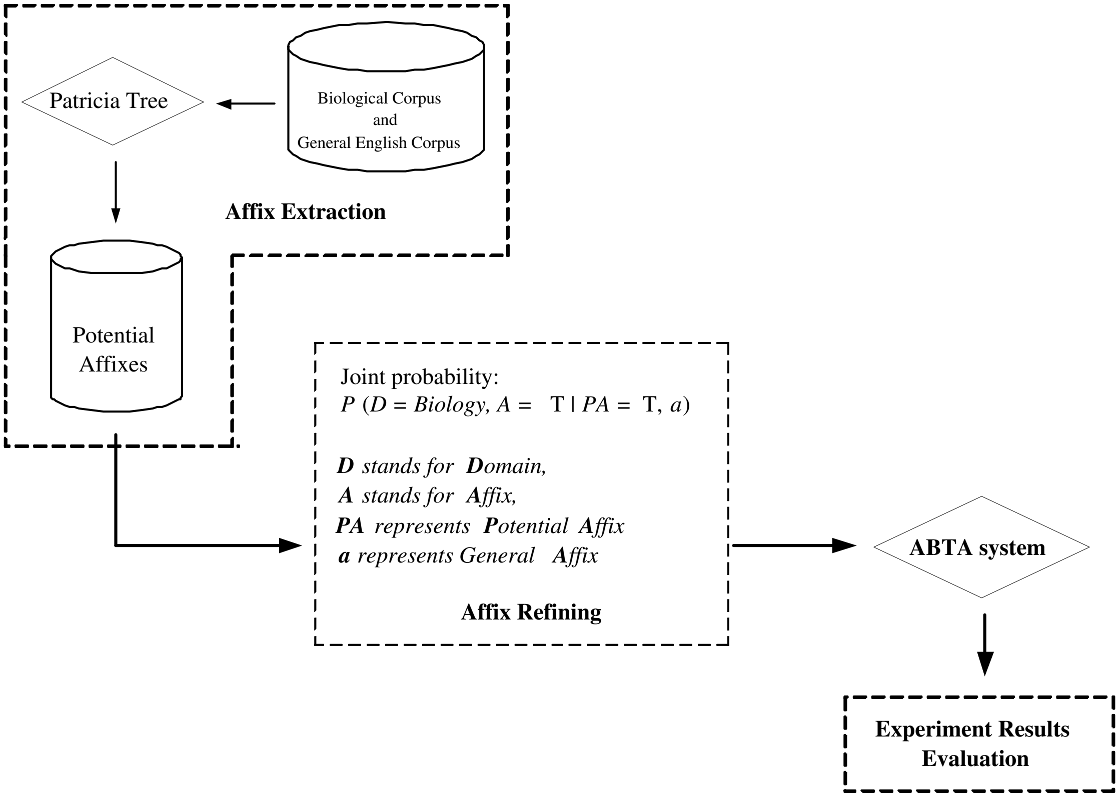

Figure 2: Experiment Design Diagram

In the affix extraction process, we first populate PATRICIA tree using all words in the combined corpus*(CC)* of a Biological Corpus *(BC)* and a General English Corpus *(GEC)*. 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 the value of its corresponding joint probability of being an English affix and a biology related affix, $P(D = \mathit{Biology}, A = \top \mid \mathit{PA} = \top, p)$. Finally, the obtained probability value of each potential affix is used to parametrize a word beginning or ending with *PA* in *BC*.

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 probability values as affix feature: a value for its potential prefix and a value for its potential suffix. We then replace the original affix feature of ABTA system [5] with our obtained joint probability values, and investigate whether these new affix information leads to equivalent or better term annotation on *BC*.

Results and Evaluation

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) [12] and Brown corpus [16]. 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 [17], 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 at the CS Faculty of Dalhousie University. The server type is SunOS sparc SUNW, Sun-Fire-880. Our experiments were mainly implemented using Perl and Python.

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 $P(A = \top \mid \mathit{PA} = \top, p) \geq 0.5$. 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.

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

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

In order to investigate whether the extracted affixes improve 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 remaining 200 abstracts are used as testing set. The evaluation measures are precision, recall and F-score. C4.5 decision tree classifier [18] is reported as the most efficient classifier which leads to the best performance among all the classifiers experimented in [5]. Therefore, C4.5 is used as the main classifier in our experiments. WEKA 3.4 machine learning toolkit [19] is applied to perform the classification algorithm. 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. 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-word biological terms, many middle words are seemingly unrelated to biology domain while many ending words directly indicate their identity, for instance, “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

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

to denote Joint Probability Values. 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. Meanwhile, the execution time of the experiments generally reduces to around 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 additional 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.

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

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.

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

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 [12] 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%.

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 of 10 obtained classification accuracies. Figure 3 was drawn to illustrate the distribution of mean classification accuracy of each training set size for both methods, with the incremental proportion of training data. The X axis in Figure 3 has been log-scaled with base 10 in order to better compare the results of two methods. It is expected that larger training set size leads to better classification accuracy of experiments. The change patterns of mean classification accuracy obtained by our method and the original ABTA system are similar. At the beginning, the mean classification accuracy increases fast and sharply with the increase of the proportion of training data from 0.25% to 10%, and then tends to keep its continuous increase at a slower rate. It is also seen that our method achieves better classification performance when the proportion of training data is under 2.5%.

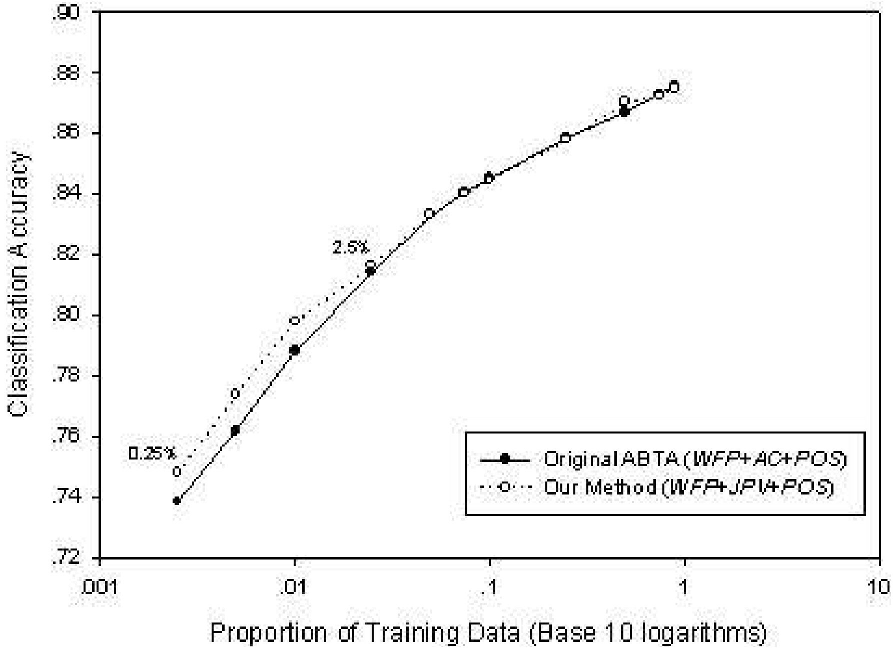

Figure 3: Distribution of Mean Classification Accuracy

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 [20] on the classification results with our hypothesis that the mean classification accuracy by our proposed method is higher than the mean classification accuracy by the 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% since the *P* values are much lower than α. Table 5 shows the *P* values of t-Test results for the training set sizes 0.25%, 0.5%, 1% and 2.5%. This demonstrates that the ABTA system adopting our method outperforms the original ABTA system in classification accuracy with smaller proportion of training data, and achieves comparable classification performance with the original ABTA system when the proportion continuouly increases. A pre-annotated data set is not always available for some domains, for example biological domain. Therefore, it is necessary for domain experts to manually annotate the raw data sets. Instead of annotating a large data set by hand, it is understandable that domain experts are more willing to manually evaluate small amount of data. The annotated data can then be used as a seed training data set for the annotation system to further annotate the experimental data. Hence, our method would be more preferable in this case as it performs better with small amount of training data.

One-tailed t-Test	Training set size			
	0.25%	0.5%	1%	2.5%
P value	0.029779	0.000605	0.000201	0.022885

Table 5: One-tailed t-Test Results

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. 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 demonstrates that the affix information acheived by the proposed method is important to accurately locate biological terms.

We further explored the ability of our method to learn from small seed training data 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 better performance than the original ABTA system when the training corpus is small. The main result of this work is 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. In the evaluation process of our experiments, for example, we integrated all the obtained joint probability values of potential affixes with every word in Genia corpus. 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.

Authors contributions

HL participated in the design of the study, performed the experiments and drafted the manuscript. CB and VK participated in the design and coordination of the study, and helped to draft and revise the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank MITACS internship and the Kanayo company, who sponsored our project.

References

1. Fukuda K, Tsunoda T, Tamura A, Takagi T: **Toward information extraction: Identifying protein names from biological papers**. In *the Pacific Symposium on Biocomputing* 1998:707–718.
2. Franzn K, Eriksson G, Olsson F, Lidn LAP, Cster J: **Protein names and how to find them**. *International Journal of Medical Informatics special issue on Natural Language Processing in Biomedical Applications* 2002, :49–61.
3. Zhou G, Su J: **Exploring deep knowledge resources in biomedical name recognition**. In *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)* 2004.
4. Jiampojamarn S, Cercone N, Kešelj V: **Automatic Biological Term Annotation Using N-gram and Classification Models**. *Master's thesis*, Faculty of Computer Science, Dalhousie University 2005.
5. Jiampojamarn S, Cercone N, Kešelj V: **Biological Named Entity Recognition using N-grams and Classification Methods**. In *the Conference Pacific Association for Computational Linguistics, PACLING'05*, Tokyo, Japan August 2005.
6. Cavnar WB, Trenkle JM: **N-Gram-Based Text Categorization**. In *Proc. SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, USA 1994:161–175.
7. Finkel J, Dingare S, Nguyen H, Nissim M, Sinclair G, Manning C: **Exploiting context for biomedical entity recognition: From syntax to the web**. In *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)* 2004.
8. Settles B: **Biomedical named entity recognition using conditional random fields and novel feature sets**. In *the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)* 2004.
9. Nenadic G, Ananiadou S: **Mining semantically related terms from biomedical literature**. *ACM Transactions on Asian Language Information Processing (TALIP)* 2006, **5**:22–43.
10. Cohen AM, Hersh WR: **A survey of current work in biomedical text mining**. *Briefings in Bioinformatics* 2005, **5**:57–71.
11. Franzn K, Eriksson G, Olsson F, Lidn LAP, Cster J: **Mining the Biomedical Literature in the Genomic Era: An Overview**. *Journal of Computational Biology* 2003, **10**(6):821–855.
12. GENIA: **GENIA Corpus, Tsujii laboratory of the University of Tokyo**. [Accessed in August 2007, http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/].
13. Shen D, Zhang J, Zhou G, Su J, Tan CL: **Effective adaptation of a Hidden Markov Model-based named entity recognizer for biomedical domain**. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, Morristown, NJ, USA: Association for Computational Linguistics 2003:49–56.
14. Lee KJ, Hwang YS, Rim HC: **Two-phase biomedical NE recognition based on SVMs**. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine*, Morristown, NJ, USA: Association for Computational Linguistics 2003:33–40.

15. Morrison DR: **PATRICIA — Practical Algorithm To Retrieve Information Coded in Alphanumeric**. *Journal of the ACM* October 1968, **15**(4):514–534.

16. Brown: **The Brown Corpus, Brown University**. [Accessed in August 2007, http://clwww.essex.ac.uk/w3c/corpus_ling/].

17. MEDLINE: **National Library of Medicine**. [Accessed in August 2007, http://www.ncbi.nlm.nih.gov/PubMed/].

18. Quinlan JR: *C4.5: programs for machine learning*. Morgan Kaufmann 1993.

19. **Weka: Machine learning software in Java**. [Accessed in August 2007, http://www.cs.waikato.ac.nz/~ml/weka/].

20. Alpaydin E: *Introduction to Machine Learning*. MIT Press 2004.