Identifying Interaction Sentences from Biological Literature Using Automatically Extracted Patterns

Haibin Liu
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada
haibin@cs.dal.ca

Christian Blouin
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada
cblouin@cs.dal.ca

Vlado Kešelj
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada
vlado@cs.dal.ca

Abstract

An important task in information retrieval is to identify sentences that contain important relationships between key concepts. In this work, we propose a novel approach to automatically extract sentence patterns that contain interactions involving concepts of molecular biology. A pattern is defined in this work as a sequence of specialized Part-of-Speech (POS) tags that capture the structure of key sentences in the scientific literature. Each candidate sentence for the classification task is encoded as a POS array and then aligned to a collection of pre-extracted patterns. The quality of the alignment is expressed as a pairwise alignment score. The most innovative component of this work is the use of a Genetic Algorithm (GA) to maximize the classification performance of the alignment scoring scheme. The system achieves an F-score of 0.834 in identifying sentences which describe interactions between biological entities. This performance is mostly affected by the quality of the preprocessing steps such as term identification and POS tagging.

1 Introduction

Recent research in information extraction (IE) in biological science has focused on extracting information about interactions between biological entities from research communications. The type of interaction of interest includes protein-protein, protein-DNA, gene regulations and other interactions between macromolecules. This work broadens the definition of the term "interaction" to include other types of concepts that are semantically related to cellular components and processes. This contrasts with the past efforts focusing strictly on molecular interactions (Blaschke et al., 1999; Ono et al., 2001). We anticipate that identifying the relationships between concepts of molecular biology will facilitate the building of knowledge models, improve the sensitivity of IE tasks and ultimately facilitate the formulation of new hypothesis by experimentalists.

The extraction of interactions is based on the heuristic premise that interacting concepts co-occur within a given section of text. The challenge is that co-occurrence certainly does not guarantee that a passage contains an interaction(Jang et al., 2006; Skusa et al., 2005). Co-occurrence is highly dependent on the definition of the section of text within which the target terms are expected to be found. A thorough comparison on the prediction of protein-protein interaction between abstract-level co-occurrence and sentence-level co-occurrence was undertaken (Raychaudhuri, 2006). It is demonstrated that abstract co-occurrence is more sensitive but less specific for interactions. At the cost of wide coverage, sentence co-occurrence increases the accuracy of interaction prediction. Since the ultimate goal of IE is to extract knowledge and accuracy is the most important aspect in evaluating the performance of such systems, it makes sense to focus the effort in seeking interaction sentences rather than passages or abstracts. Not every co-occurrence in sentences implies a relationship that expresses a fact. In the 2005 Genomics Track dataset, 50% of all sentence co-occurrences of entities correspond to definite relationships while the rest of the co-occurrences only convey some possible relationships or contain no relationship of interest (Li et al., 2005). Therefore, more sophisticated text mining strategies are required to classify sentences that describe interactions between co-occurring concepts.

In the BioCreative II challenge [1], teams were asked to determine whether a given passage of text contained information about the interaction between two proteins. This classification task worked at the abstract level and the interacting protein pairs were not required to be extracted. The task for the Learning Language in Logic

[1] http://biocreative.sourceforge.net/

Proceedings of the Workshop on BioNLP, pages 133–141,
Boulder, Colorado, June 2009.

(LLL'05) challenge [2] was to build systems that extract interactions between genes or proteins from biological literature. From individual sentences annotated with agent-target relations, patterns or models had to be learned to extract these interactions. The task focused on extracting only the interacting partners. The context of an interaction may also be critical to the validity of the extracted knowledge since not all statements found in the literature are always true.

In this work, we propose an approach to automatically extract patterns containing relevant interaction between biological concepts. This extraction is based on the assumption that biological interactions are articulated by a limited number of POS patterns embedded in sentences where entities/concepts are co-occurring. The extracted patterns are then applied to identify interaction sentences which describe interactions between biological entities. Our work aims to identify precise sentences rather than passages. Because of the nature of the patterns, we hope that some of the contextual information present in interaction sentences also play a role in the classification task.

The rest of the paper is organized as follows: In Section 2, we review recent research advances in extracting biological interactions. Section 3 describes an experimental system designed for our work. Sections 4, 5 and 6 elaborate the approaches and algorithms. Performance is evaluated in Section 7. Finally, Section 8 summarizes the paper and introduces future work.

2 Related work

Early on, Blaschke (Blaschke et al., 1999) employed patterns to predict the presence of a protein-protein interaction. A series of patterns was developed manually to cover the most obvious descriptions of protein functions. This process was based on a set of keywords, including interaction verbs, that are commonly used to describe this type of interaction. A sentence extraction system BioIE (Divoli and Attwood, 2005) also uses patterns to extract entire sentences related to protein families, protein structures, functions and diseases. The patterns were manually defined and consisted of single words, word pairs, and small phrases.

Although systems relying on hand-coded patterns have achieved some success in extracting biological interactions, the strict requirement of dedicated expert work is problematic. Moreover, each type of interaction may require a definition of many different patterns including different arrangements and different variants of the same keyword. Manually encoding all patterns encountered in a corpus is time-consuming and potentially impractical in real applications. Thus, automatically learning such patterns is an attractive solution.

An approach which combines dynamic programming and sequence alignment algorithms as normally used for the comparison between nucleotide sequences was introduced by Huang *et al.* (Huang et al., 2004). This approach is designed to generate patterns useful for extracting protein-protein interactions. The main problem with this approach is that the scoring scheme that is required to implement the alignment algorithm is difficult to define and contains a potentially large number of free parameters. We propose a method based on Genetic Algorithm (GA) heuristics to maximize the alignment procedure for the purpose of classification. GAs were also used as a learning strategy to train finite state automata for finding biological relation patterns in texts(Plake et al., 2005). It was reported (Bunescu et al., 2005; Hakenberg et al., 2005) that automatically learned patterns identify biological interactions even more accurately than hand-coded patterns.

3 Overview of system design

In this work, we have designed an experimental system to facilitate the automatic extraction of biological interaction patterns and the identification of interaction sentences. It consists of three major modules: biological text preprocessing, interaction pattern extraction, and interaction sentence identification.

Biological text preprocessing reformats the original biological texts into candidate sentences. A pattern learning method is then proposed to automatically extract the representative patterns of biological interactions. The obtained patterns are further used to identify instances that evidently describe biological interactions. Poor performance during preprocessing will have detrimental effects on later stages. In the following sections, we will describe each component.

4 Biological text preprocessing

4.1 Sentence preparation

A heuristic method is implemented to detect sentence boundaries (Mikheev, 2002) based on the assumption that sentences are usually demarcated by some indicative delimiting punctuation marks in order to segment the biological texts into sentence units. Captions and headings that are not grammatically valid sentences are therefore detected and further eliminated for our work.

[2] http://genome.jouy.inra.fr/texte/LLLchallenge/

4.2 Part-of-Speech tagging

POS tagging is then performed to associate each word in a sentence with its most likely POS tag. Because subsequent processing steps typically depend on the tagger's output, high performance at this level is crucial for success in later stages. A statistical tagger Lingua::EN::Tagger [3] is used to perform this task.

4.3 Biological term annotation

A learning-based biological term annotation system, ABTA (Jiampojamarn et al., 2005), is embedded in our system. The type of terms includes molecules, such as genes, proteins and cell lines, and also biological processes. Examples of biological processes as entities are: "T cell activation" and "IL-2 gene transcription". We consider that a broader definition of biological term will include more facts from literature, thus leading to more general use of interaction patterns for IE tasks.

ABTA considers the longest expression and ignores embedded entities. Further, instead of distinguishing terms from their relevant biology concepts, a unified tag "*BIO*" is assigned to all the identified terms. We aim to discover patterns of the general interactions between biological concepts, not only the interactions between molecules, e.g., protein-protein interaction.

Tags like *NN*(noun) and *VB*(verb) are typically used to define entities and the action type of interactions, and thus they are indispensable. However, tags such as *JJ*(adjective) and *RB*(adverb) could occur at different positions in a sentence. We decided to remove these tags to prevent the combinatorial effect that these would induce within the set of extracted patterns.

4.4 Text chunking

Next, a rule-based text chunker (Ramshaw and Marcus, 1995) is applied on the tagged sentences to further identify phrasal units, such as base noun phrases *NP* and verbal units *VB*. This allows us to focus on the holistic structure of each sentence. Text chunking is not applied on the identified biological terms. In order to achieve more generalized interaction patterns, a unified tag "*VB*" is used to represent every verbal unit instead of employing different tags for various tenses of verbs.

As a result of preprocessing, every sentence is represented by its generalized form as a sequence of corresponding tags consisting of POS tags and predefined tags. Table 1 summarizes the main tags in the system.

A biological interaction tends to involve at least three objects: a pair of co-occurring biological entities connected by a verb which specifies the action type of the interaction. Thus, a constraint is applied that only sentences satisfying form "BioEntity A – Verb – BioEntity B" will be preserved as candidate sentences to be further processed in the system. It is possible that the presence of two entities in different sentence structures implies a relationship. However, this work assumes the underlying co-occurrence of two concepts and a verb in the interest of improving the classification accuracy.

Tag name	Tag description	Tag type
BIO	Biological entity	Predefined
NP	Base noun phrase	Predefined
VB	Verbal unit	Predefined
IN	Preposition	POS
CC	Coordinating conjunction	POS
TO	to	POS
PPC	Punctuation comma	POS
PRP	Possessive 2nd determiner	POS
DET	Determiner	POS
POS	Possessive	POS

Table 1: Main tags used in the system

The obtained candidate sentences are split into training and testing sets. The training set is used to extract the representative patterns of biological interactions. The testing set is prepared for identifying sentences that evidently describe biological interactions.

5 Interaction pattern extraction

5.1 PATRICIA trees

The method we propose to extract interaction patterns from candidate sentences is based on the use of PATRICIA trees (Morrison, 1968). A 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). It has been applied to many large information retrieval problems (Chien, 1997; Chen et al., 1998).

In our work, a PATRICIA tree is used for the first time to facilitate the automatic extraction of interaction patterns. All training sentences are inserted and stored in a generic PATRICIA tree from which the common patterns of POS tags can be efficiently stored and the tree structure used to compute relevant usage statistics.

5.2 Potential pattern extraction

Patterns of straightforward biological interactions are frequently encountered in a range of actual sentences. Conversely, vague relationships or complex interactions patterns are seldom repeated. Therefore, the

[3] http://search.cpan.org/~acoburn

premise of this work is that there is a set of frequently occurring interaction patterns that matches a majority of stated facts about molecular biology. In this work, a *biological interaction pattern* is defined as follows:

Definition 5.1. A biological interaction pattern bip is a sequence of tags defined in Table 1 that captures an aggregate view of the description of certain types of biological interactions based on the consistently repeated occurrences of this sequence of tags in different interaction sentences. $BIP = \{bip_1, bip_2, \cdots, bip_k\}$ represents the set of biological interaction patterns.

We first extract potential interaction patterns by populating a PATRICIA tree using training sentences. Every node in the tree contains one or more system tags, which is the preceding tag sequence of its descendant nodes in each sentence. Every sentence is composed of a path of system tags from the root to a leaf. Hence, we propose that the sequence of system tags that can be formed from traversing the nodes of the tree is a potential pattern of biological interactions. At the same time, the occurrence frequency of each pattern is also retrieved from the traversal of tree nodes.

A predefined frequency threshold f_{min} is used as a constraint to filter out patterns that occur less than f_{min} times. It has been demonstrated that if an interaction is well recognized, it will be consistently repeated (Blaschke et al., 1999; Ono et al., 2001). The generalization and the usability of patterns can be controlled by tuning f_{min}. Further, some filtering rules are adapted to control the form of a pattern and enhance the quality of the discovered patterns, such as if a pattern ends with a tag *IN, VB, CC* or *TO*, the pattern will be rejected. Flexibility in setting this threshold can be applied to meet special demands. Algorithm 1 shows our pattern learning method which has a time complexity of $O(n)$ in the size of candidate sentences, n.

Algorithm 1 Patricia-Tree-based Extraction of Biological Interaction Patterns

Input: Candidate Sentences $CS \in$ Biological text; a predefined threshold f_{min}; a set of filtering rules FR
Output: BIP : Set of biological interaction patterns

```
BIP ← ∅;  PT ← ∅                          //PT : Patricia Trie
for all sentences s ∈ CS do
    PT ← Insert(s)                        //Populating Patricia Tree
for all nodes n_i ∈ PT do
    bip_i ← Pattern(n_i)                  //Concatenating tags in nodes
    from root to n_i, which is a potential pattern
    if Count(bip_i) ≥ f_min and bip_i does not meet FR
    then
        //Count(bip_i) returns No. of occurrences of bip_i;
        BIP ← bip_i
```

5.3 Interaction verb mining

Although the obtained patterns are derived from the candidate sentences possessing the form "BioEntity A – Verb – BioEntity B", some of them may not contain facts about biological interactions. This is possible if the action verbs do not describe an interaction. Quite a few verbs, such as "report", "believe", and "discover", only serve a narrative discourse purpose. Therefore, mining the correct interaction verbs becomes an important step in the automatic discovery of patterns. We decided to perform the method applied in (Huang et al., 2004) to mine a list of interaction verbs. This will be used to further improve the relevance of achieved patterns by filtering out patterns formed by the sentences in which the action verbs are not on the list.

6 Interaction sentence identification

Once the biological interaction patterns are obtained, we perform interaction sentence identification on testing sentences. For our work, they are partitioned into two sets: interaction sentences which explicitly discuss interactions between entities, and non-interaction sentences which do not describe interactions, or merely imply some vague relationships between entities. The task of interaction sentence identification is treated as a classification problem to differentiate between interaction sentences and non-interaction sentences.

6.1 Pattern matching scoring

We first perform pattern matching by iteratively applying the interaction patterns to each testing sentence. This is done using sequence alignment which calculates the degree of the similarity of a sentence to an interaction pattern. Since patterns capture various ways of expressing interactions among sentences, a high similarity between an interaction sentence and a pattern is expected. Therefore, we conjecture that the alignment scores can be used to discriminate some type of interaction sentences from other types of sentences.

The scoring scheme involved in the pattern matching consists of penalties for introducing gaps, match rewards and mismatch penalties for different system tag pairs. Table 2 presents an example scoring scheme for main tags. Penalties and rewards are denoted respectively by negative and positive values.

As a variation of global alignment, an end-space free alignment algorithm is implemented to facilitate the alignment between patterns and testing sentences. The shortest pattern is always preferred for a sentence in case that same alignment score is achieved by multiple

Tag	Gap	Match	Mismatch
BIO	-10	+8	-3
NP	-8	+6	-3
VB	-7	+7	-3
IN	-6	+5	-1
CC	-6	+5	-1
TO	-1	+5	-1
PPC	-1	+3	-1
PRP	-1	+3	-1
DET	-1	+3	-1
POS	-1	+3	-1

Table 2: An alignment scoring scheme for system tags

patterns. As a result, each sentence is assigned to its most appropriate pattern along with a maximum alignment score. Therefore, an interaction sentence will be highlighted with a high alignment score by its most similar interaction pattern, while a non-interaction sentence will be characterized by a low alignment score indicating rejections by all patterns. Essentially, this procedure can be seen as a variation of the well-known k Nearest Neighbors classification method, with $k = 1$.

6.2 Performance evaluation

We then evaluate whether the alignment scores can be used to classify the testing sentences. We have proposed two independent evaluation measures: statistical analysis (SA) and classification accuracy (AC).

SA measures whether the scoring difference between the mean of interaction sentences and the mean of non-interaction sentences is statistically significant. If the difference is significant, there will be a tendency that interaction sentences outscore non-interaction sentences in alignment. Hence, it would be reliable to use alignment scores to classify testing sentences. Although non-interaction sentences could come from the same documents as interaction sentences and discuss concepts that are associated with the target interactions, we assume that interaction sentences and non-interaction sentences are two independent samples.

The statistical two-sample z test (Freund and Perles, 2006) is performed with the null hypothesis that there is no scoring difference between the means of interaction and non-interaction sentences. A comparatively large z will lead to the rejection of the null hypothesis. Naturally, the increase of z value will increase the difference between the means and therefore conceptually keep pushing the overall scoring distributions of two samples further away from each other. Consequently, interaction sentences can be separated from non-interaction sentences according to alignment scores. In reality, the distinction between interaction and non-interaction sentences is not absolute. Thus, the scoring distributions of two samples can only be distanced by a certain maximum value of z depending on the scoring scheme applied in pattern matching.

Conversely, AC measures the proportion of correctly classified testing sentences, including both interaction and non-interaction sentences, to the total testing sentences. An appropriate threshold T is determined for obtained alignment scores to differentiate between interaction and non-interaction sentences, and to facilitate the calculation of classification accuracy.

It is not possible to evaluate the performance without correctly pre-labeled testing sentences. We decided to manually classify the testing sentences in advance by assigning each sentence an appropriate label of interaction or non-interaction. This work was done by two independent experts, both with Ph.D. degrees in molecular biology or a related discipline.

6.3 Scoring scheme optimization

The scoring scheme applied in pattern matching has a crucial impact on the performance of interaction sentence identification. An interesting problem is whether there exists an optimal scoring scheme covering the costs of gap, match and mismatch for different system tags in the pattern matching alignment, which is destined to achieve the best performance on classifying testing sentences. To the best of our knowledge, no efforts have been made to investigate this problem. Instead, an empirical or arbitrary scoring scheme was adopted in previous research for the pairwise alignments (Huang et al., 2004; Hakenberg et al., 2005). We have proved that the problem is NP-hard by reducing a well-known NP-hard problem 3-SAT to this problem. The proof is not presented in this work.

A genetic algorithm (GA) is used as a heuristic method to optimize parameters of the scoring scheme for sentence classification. The costs of penalties and rewards for different system tags are encoded by integer values within two predefined ranges: [-50, 0) and (0, 50], and assembled as a potential solution of scoring scheme, which consists of 30 parameters covering the costs for tags in the alignment as listed in Table 2. The two evaluation measures SA and AC are used as the fitness function for GA respectively with the goal of maximizing z value or classification accuracy.

GA is set up to evolve for 100 generations, each of which consists of a population of 100 potential solutions of scoring scheme. GA starts with a randomly

generated population of 100 potential solutions and proceeds until 100 generations are reached. The number of generations and the population size are decided with consideration of the runtime cost of evaluating the fitness function, which requires running the scoring algorithm with each sentence. A large number of generations or a large population size would incur an expensive runtime cost of evaluation.

In addition, we further divide the labeled set of candidate sentences into two subsets: The first dataset is used to optimize parameters of the scoring scheme, while the second dataset, testing set, is used to test the achieved scheme on the task of sentence classification.

7 Results and evaluation

7.1 Dataset

Our experiments have been conducted on Genia corpus (v3.02) [4], the largest, publicly available corpus in molecular biology domain. It consists of 2,000 biological research paper abstracts and is intended to cover biological reactions concerning transcription factors in human blood cells. The information of sentence segmentation, word tokenization, POS tagging and biological term annotation is also encoded in the corpus.

7.2 Biological text preprocessing results

Evaluated using the inherently equipped annotation information, our system achieves nearly 99% accuracy on segmenting sentences. Further, it obtains an overall POS tagging accuracy of 91.0% on 364,208 individual words. We noticed that the tagging information encoded in Genia corpus is not always consistent throughout the whole corpus, thus introducing detrimental effects on the tagging performance. Also, considering that the tagger is parameterized according to the general English domain, porting this tagger to the biology domain is accompanied by some loss in performance.

The system reaches an F-score of 0.705 on annotating all biological terms including both multi-word and single word terms. After performing text chunking, the system produces a set of candidate sentences. We further perform text chunking on Genia corpus based on its encoded annotations and use the resulting set of sentences for the subsequent experiments to provide a gold standard to which results produced based on our system annotations can be compared. Table 3 presents some statistics of the preprocessed dataset. For each type of annotations, we randomized the candidate sentence set and chose 12,525 candidate sentences as the training set to extract biological interaction patterns. The rest of candidate sentences are prepared as the testing set.

Attributes	Genia	Our system
Total preprocessed sentences	18,545	18,355
Candidate sentences	16,272	17,525
Training set sentences	12,525	12,525
Testing set sentences	6,020	5,000

Table 3: Statistics of experimental dataset

7.3 Interaction pattern extraction results

$f_{min} = 5$ is used to filter out the potential patterns that appear less than 5 times in the training set. Evaluated by domain experts, lists of 300 interaction verbs and 700 non-interaction verbs are obtained from 12,525 training sentences with Genia annotations. Inflectional variants of the verbs are also added into the lists.

Refined by the filtering rules and the interaction verbs, a final set of representative patterns of biological interactions are obtained from Algorithm 1. We performed our proposed pattern learning method on training sentences of both the GENIA and our own annotations. There are respectively 241 and 329 potential patterns. Of these, 209 and 302 were extracted. Interestingly, only 97 extracted patterns are common to both annotation schemes.

Table 4 lists the 10 most frequent interaction patterns based on Genia annotations. For instance, a training sentence conforming to the second pattern is "The expression of the QR gene is regulated by the transcription factor AP-1." (MEDLINE: 96146856).

Pattern count	Pattern
264	*BIO VB BIO IN BIO*
261	*NP IN BIO VB IN BIO*
182	*NP IN BIO VB BIO*
162	*BIO IN BIO VB IN BIO*
160	*BIO VB IN BIO IN BIO*
143	*NP IN BIO VB IN NP IN BIO*
142	*NP VB IN BIO VB BIO*
138	*PRP VB IN BIO VB BIO*
126	*BIO VB NP IN BIO IN BIO*
121	*NP IN BIO VB NP IN BIO*

Table 4: Extracted Biological Interaction Patterns

7.4 Interaction sentence identification results

Since the total testing sentence set is large, we decided to randomly extract 400 sentences from it as the sample set for our task. The 400 sentences were manu-

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

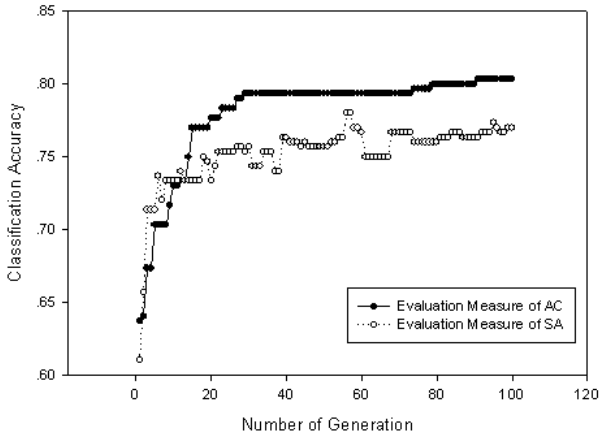

Figure 1: AC comparison between two measures

ally pre-labeled into two classes: interaction and non-interaction. Further, a subset of 300 testing sentences was used by GA to optimize parameters of the scoring scheme, while the remaining 100 sentences were prepared to test the achieved scheme on sentence classification. The distribution of class labels of the sample sentences is shown in Table 5.

Class label	300 sentences		100 sentences	
	No.	%	No.	%
Interaction	158	52.67	53	53
Non-interaction	142	47.33	47	47

Table 5: Class distribution of sample sentences

7.4.1 Comparison between two measures

We applied the evaluation measures, SA and AC, respectively to the subset of 300 testing sentences as the fitness function for GA, and recorded the scoring scheme of every generation resulted from GA. Figure 1 presents the distribution of achieved classification accuracy in terms of each scoring scheme optimized by GA. This comparison is done with respect to the generation and evaluated on 300 testing sentences using the annotations from the Genia corpus.

The achieved classification accuracy for AC generally outperforms the classification accuracy derived by SA. It reaches its highest classification accuracy 80.33% from the 91th generation. Therefore, AC is considered more efficient with the system and becomes our final choice of fitness function for GA.

7.4.2 Results of sentence identification

GA results in an optimized performance on the 300 sentences. It also results in an optimized scoring scheme along with its associated scoring threshold T, which are then applied together to the other 100 testing sentences. Table 6 and 7 present the system performance on the two sets respectively to both annotations.

Experimental Results	Genia		Our system	
	Interaction	Non	Interaction	Non
Precision	0.757	0.887	0.704	0.702
Recall	0.928	0.665	0.761	0.640
F-score	0.834	0.750	0.731	0.670
Overall $AC(\%)$	80.33		70.33	

Table 6: Performance on 300 testing sentences

Experimental Results	Genia		Our system	
	Interaction	Non	Interaction	Non
Precision	0.739	0.762	0.676	0.697
Recall	0.792	0.723	0.755	0.638
F-score	0.765	0.742	0.713	0.666
Overall $AC(\%)$	75.96		70.00	

Table 7: Performance on 100 testing sentences

Table 6 shows that when using the Genia annotations the system achieves an 0.834 F-score in identifying interaction sentences and an overall AC of 80.33%, which is much higher than the proportion of either interaction or non-interaction sentences in the 300 sentence subset. This indicates that the system performs well on both classes. In 100 generations GA is not able to evolve a scoring scheme that leads to an AC above 80.33%. Moreover, our system annotations achieve a lower performance than Genia annotations. We attribute the difference to the accuracy loss of our system annotations in the preprocessing steps as inaccurate annotations will lead to inappropriate patterns, thus harming the performance of sentence identification. For Genia annotations, the performance on the 100 testing sentences suggests an overfitting problem.

There are a number of preprocessing steps that affect the final classification performance. However, even assuming an ideal preprocessing of the unstructured text, our method relies on the assumption that all interaction sentences are articulated by a set of POS patterns that are distinct to all other types of sentences. The manual annotation of the training/testing set was a difficult task, so it is reasonable to assume that this will also be difficult for the classifier. The use of passive voice and the common use of comma splicing within patterns makes sentence-level classification an especially difficult task. Another source of interactions that our system cannot identify are implied and assume a deeper semantic understanding of the concepts them-

selves. Other sentences are long enough that the interaction itself is merely a secondary purpose to another idea. All of these factors pose interesting challenges for future development of this work.

Moreover, we also experimented with 10 empirical scoring schemes derived from previous experiments on the 300 sentences respectively, including the scheme in the Table 2. Several fixed thresholds were attempted for obtained alignment scores to differentiate between interaction and non-interaction sentences. Without using GA to optimize parameters of the scoring scheme, the best performance of 10 empirical schemes is an overall AC of 65.67%, which is outperformed at the 3rd generation of the GA optimization with Genia annotations.

7.5 System performance comparison

Within the framework of our system, we further conducted experiments on the same dataset for sentence identification using interaction patterns generated by another pattern generating algorithm (PGA) (Huang et al., 2004) in order to compare with the performance of patterns obtained by our pattern learning method.

In our implementation, PGA iterates over all pairs of candidate sentences in the training set and calculates the best alignment for each pair in terms of the cost scheme of gap penalties proposed (Huang et al., 2004). Each consensus sequence from the optimal alignment of each pair forms a pattern. The filter rules proposed are also applied. PGA has a time complexity of $O(n^2)$ in the size of candidate sentences, n. Hence, our proposed pattern learning method is much more efficient when dealing with large collections of biological texts. PGA produces a large number of patterns, even with $f_{min} = 5$ and other filtering criteria. There are 37,319 common patterns between two types of annotations.

Attributes	Genia	Our system
Potential patterns ($f_{min} = 5$)	476,600	387,302
Extracted patterns ($f_{min} = 5$)	176,082	88,800

Table 8: Pattern extraction results of PGA

In order to make a direct comparison, we decided to experiment with the same number of interaction patterns. For Genia annotations, we chose the most frequent 209 patterns generated by PGA to compare with the 209 patterns by our method. For our system annotations, two sets of 302 patterns are employed. Further, it is found that there are 96 common patterns between the two sets of 209 patterns for Genia annotations, and 153 common patterns between the two sets of 302 patterns for our system annotations. Table 9 and 10 present the results of sentence identification of PGA. The results show that patterns generated by PGA do not perform as well as patterns obtained by our method.

Experimental	Genia		Our system	
Results	Interaction	Non	Interaction	Non
Precision	0.721	0.869	0.663	0.699
Recall	0.918	0.606	0.785	0.556
F-score	0.808	0.714	0.719	0.619
Overall $AC(\%)$	77.00		67.67	

Table 9: Performance of PGA on 300 testing sentences

Experimental	Genia		Our system	
Results	Interaction	Non	Interaction	Non
Precision	0.664	0.796	0.698	0.635
Recall	0.849	0.574	0.566	0.766
F-score	0.745	0.667	0.625	0.694
Overall $AC(\%)$	71.98		66.00	

Table 10: Performance of PGA on 100 testing sentences

8 Conclusion and future work

In this paper, a novel approach is presented to automatically extract the representative patterns of biological interactions, which are used to detect sentences that describe biological interactions. We conducted the experiments on our designed system based on the Genia corpus. By means of a genetic algorithm, the system achieves an 0.834 F-score using Genia annotations and an 0.731 F-score using our system annotations in identifying interaction sentences by evaluating 300 sentences. By applying the optimized scoring scheme to another set of 100 sentences, the system achieves comparable results for both types of annotations. Furthermore, by comparing with another pattern generating algorithm, we infer that our proposed method is more efficient in producing patterns to identify interaction sentences.

In our future work, we would like to employ the obtained interaction patterns to guide the extraction of specific interactions. The matching between patterns and sentences will be performed and the matched parts of each sentence will be extracted as candidate interactions. Further reasoning processes can be performed by means of available biological ontologies, such as UMLS Semantic Network (Mccray and Bodenreider, 2002) and Gene Ontology (Consortium, 2001), to infer new relations from the initial interactions. Such processes can be employed to derive additional biological knowledge from existing knowledge, or test for biological consistency of the newly entered data.

References

Christian Blaschke, Miguel A. Andrade, Christos Ouzounis, and Alfonso Valencia. 1999. Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 60–67. AAAI Press.

Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk W Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Keh-Jiann Chen, Wen Tsuei, and Lee-Feng Chien. 1998. Pat-trees with the deletion function as the learning device for linguistic patterns. In *Proceedings of the 17th international conference on Computational linguistics*, pages 244–250, Morristown, NJ, USA. Association for Computational Linguistics.

Lee-Feng Chien. 1997. Pat-tree-based keyword extraction for chinese information retrieval. *SIGIR Forum*, 31(SI):50–58.

Gene Ontology Consortium. 2001. Creating the gene ontology resource: design and implementation. *Genome Research*, 11(8):1425–1433.

Anna Divoli and Teresa K. Attwood. 2005. Bioie: extracting informative sentences from the biomedical literature. *Bioinformatics*, 21(9):2138–2139.

John E. Freund and Benjamin M. Perles. 2006. *Modern Elementary Statistics*. Prentice Hall.

Jorg Hakenberg, Conrad Plake, Ulf Leser, Harald Kirsch, and Dietrich Rebholz-Schuhmann. 2005. Lll'05 challenge: Genic interaction extraction with alignments and finite state automata. In *Proceedings of Learning Language in Logic Workshop (LLL'05) at ICML*, page 38C45, Bonn, Germany.

Minlie Huang, Xiaoyan Zhu, Yu Hao, Donald G. Payan, Kunbin Qu, and Ming Li. 2004. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20:3604–3612.

Hyunchul Jang, Jaesoo Lim, Joon-Ho Lim, Soo-Jun Park, Kyu-Chul Lee, and Seon-Hee Park. 2006. Finding the evidence for protein-protein interactions from pubmed abstracts. *Bioinformatics*, 22(14):e220–e226.

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

Jiao Li, Xian Zhang, Yu Hao, Minlie Huang, and Xiaoyan Zhu. 2005. Learning domain-specific knowledge from context–thuir at trec2005 genomics track. In *Proceedings of 14th Text Retrireval Conference (TREC2005)*, Gaithersburg, USA.

Alexa T. Mccray and Olivier Bodenreider. 2002. A conceptual framework for the biomedical domain. In *Semantics of Relationships, Kluwer*, pages 181–198. Kluwer Academic Publishers.

Andrei Mikheev. 2002. Periods, capitalized words, etc. *Comput. Linguist.*, 28(3):289–318.

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

Toshihide Ono, Haretsugu Hishigaki, Akira Tanigami, and Toshihisa Takagi. 2001. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161.

Conrad Plake, Jorg Hakenberg, and Ulf Leser. 2005. Learning patterns for information extraction from free text. In *Proceedings of AKKD 2005*, Karlsruhe, Germany.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, Somerset, New Jersey.

Soumya Raychaudhuri. 2006. *Computational Text Analysis: For Functional Genomics and Bioinformatics*. Oxford University Press.

Andre Skusa, Alexander Ruegg, and Jacob Kohler. 2005. Extraction of biological interaction networks from scientific literature. *Brief Bioinform*, 6(3):263–276.