2011 11th IEEE International Conference on Bioinformatics and Bioengineering

Evolutional dependency parse trees for biological relation extraction

Hung-Yu Kao
Department of Computer Science and Information Engineering,
National Cheng Kung University,
Tainan, Taiwan
hykao@mail.ncku.edu.tw

Yi-Tsung Tang
Department of Computer Science and Information Engineering,
National Cheng Kung University,
Tainan, Taiwan
p7895125@mail.ncku.edu.tw

Jian-Fu Wang
Department of Computer Science and Information Engineering,
National Cheng Kung University,
Tainan, Taiwan
jfwang@ikmlab.csie.ncku.edu.tw

***Abstract*—Due to the rapid growth in biological technology, the development of high-quality information extraction systems is needed and still remains a challenge. Several recently proposed approaches to biological relation extraction are based on machine learning techniques on lexical and syntactic information. Most use the dependency path between two genes/proteins instead of the whole dependency tree of a sentence for identifying relationships. However, the dependency path may not have any node between two entities. If a limited set of annotated training corpora is used for the construction of tree information of biological relationships, the training corpus will lack some sentence structures and cannot predict whether the sentence has a biological relationship. In this paper, we developed a biological relation extraction system called Evolutional Tree Extraction System – ETree. We extended the dependency path to the dependency subtree and developed a method that can automatically expand and prune these existing dependency subtrees into various dependency subtrees. These dependency subtrees are called "Evolutional Trees" and are used to predict the biological relationship sentences.**

Keywords : text mining, gene regulation, relation extraction

I. INTRODUCTION

Due to rapid growth in biological technology, an overwhelming amount of new research results and information has been published. PubMed, includes over 19 millions of research articles from the biological domain, which makes the acquisition of new or useful information from literature difficult and time-consuming for biological scientists. The automatic and efficient extraction of information from biological literature is one of the most important tasks in the biological domain.

Biological relation extraction is the process of finding various relationships, such as gene regulatory relationships, protein-protein interactions and gene-disease relationships, between pairs of entities in texts. Many approaches to relation extraction have been applied to the biological domain. Recently proposed approaches are based not only on explicit textual information contained in biological literature but also on the syntactic information. These approaches use complex linguistic analyses, such as part of speech (POS), and the lexical and syntactic structures of the sentences. These techniques can be applied to extract a variety of different biological relationships, such as gene-disease relationships [1], molecular events [2], gene regulatory relationships [3-5] and protein-protein interactions [6-8]. Some have applied a few rules from the grammatical relationships of the sentences parsed by an NLP Parser. A previous study proposed a rule-based mechanism and applied Stanford Parser to extract biological relationships from biological texts [9]. Another study developed a relationship extraction system called RelEx [10]. This system established the chunk dependency parse trees derived from the dependency parse trees. To ensure higher accuracy, the authors proposed four relationship-filtering rules in the post-processing steps and ultimately reached their best performance on the Learning language in logic (LLL) data set of LLL-challenge 2005[11, 12]. The task of the LLL challenge 2005 is to extract biological interactions from a set of sentences. Other works have defined features to distill the information from the sentence structure, such as a dependency parse tree, which is the dependency path between two entities [3].

Many groups have used the dependency path between two entities on the parse tree for learning relationships because they consider the words between two entities to have important information for determining relationships. In addition, a previous study has developed different systems for extracting gene and protein interactions in biomedical literature [13]. They extracted the shortest dependency path between two proteins and then applied a large number of features from the path, such as the number of entities on the path, the length of the dependency path and the number of nodes on the path. Some previous studies applied robust SVMs to extract and determine the directionality of relationships trained on the GENIA corpus [14, 15]. By contrast, a system reported that uses more general structures, such as syntactic information, to construct feature vectors that can be used by SVM to decide whether a sentence has a protein-protein interaction [7]. In some previous studies, the authors proposed a relation extraction method for protein-protein interactions [6, 16]. They defined the cosine similarity and the edit-distance similarity functions among the dependency path and then applied SVM and k-nearest-neighbor to predict the protein-protein interactions.

By contrast, structural kernel-based methods are alternatives to feature-based methods and have been frequently applied to relation extraction [17-19]. In addition, the authors proposed a kernel-based method to automatically extract relationships from biomedical literature [20]. They modified the standard tree kernel by incorporating a trace kernel to capture more contextual information. Some studies modeled the syntactic structure by utilizing the convolution

978-0-7695-4391-8/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBE.2011.33

kernel over parse trees with SVM [21]. In addition, they also proposed a composite kernel, which combined the convolution tree kernel with a linear kernel for relation extraction. The composite kernel can capture both flat and structured features and easily scale to include more features with favorable results on the ACE benchmark corpora.

In addition, four kernels and machine-learning (ML) methods were proposed and applied for extracting protein/gene interactions [22]. The shortest dependency path between two protein/genes in a dependency parse tree was assumed to provide a more concise representation of information needed for relation prediction by restricting the learning features to elements inside the path. The kernel is a kind of similarity function for features derived from the shortest dependency path between two protein/genes on the dependency parse tree for a sentence. This group ultimately achieved a promising performance on the LLL data set. A previous work relies on lexical, POS and syntactic information in the dependency path between two entities [22, 23]. Besides, extracting the dependency path between two entities, the researchers proposed the all-dependency-paths kernel to identify protein/gene interactions [22]. They considered a parse tree of a sentence as a dependency graph. In addition, they also considered the dependencies outside the dependency path connecting two entities as well as dependencies on the dependency path. They assigned a weight of 0.9 to the edges of the dependency path and a weight of 0.3 to other edges. The weighting scheme can help emphasize the dependencies on the dependency path. Therefore, some relevant words outside of the dependency path can be included for relationship prediction.

In utilizing lexical and syntactic information, most of the machine learning-based approaches to biological relation extraction suggest using the dependency path between two genes or proteins instead of the whole dependency parse tree. The reason is to remove unnecessary words in a sentence that have no effect on the relationship identification. However, the dependency path between two entities may not have any nodes for relation extraction. For example in Figure 1, the dependency path between "ap1" and "epo" entities in the red dotted square only contains the entities "ap1" and "epo", but some important words such as "target" and "gene" are not on the path. In addition, if the annotated training corpus is not enough to cover all sentences of possible biological relationships, the consequence classifier or prediction system cannot work well definitely.. Therefore, the information of the dependency path is insufficient for identifying and extracting a regulatory relationship from a sentence. The incompleteness of the training corpus will lack some sentence structures and cannot predict whether the sentence has a biological relationship.

To consider the most useful information and include different sentence structures in the training set, we developed a biological relation extraction system called Evolutional Tree Extraction System (ETree). This system first extends the dependency path to the dependency subtree. The dependency subtree contains not only the path between two entities, but also contains the children nodes belonging to these two entities such as the blue square in Figure 1. For example in Figure 1, some important words belonging to the gene "epo" are "its", "target" and "gene". These nodes indicate that 'epo" is the target gene of "ap1". Thus, the structure of the dependency subtree can also be regarded as part of the sentence structure. We assume that these additional children nodes provide effective information for extracting gene regulatory relationships. In addition, the dependency paths may not cover lots of sentences structures of biological relations while the training corpus is small. A previous study proposed a syntactic pruning method in dependency graphs for the biological event extraction [24], such as regulation, protein interaction and localization. This syntactic pruning method can extract the direct relation between biological entities and verbs within a dependency relation graph. In this paper, we propose the expansion and pruning methods to expand and prune existing dependency subtrees in the original training sentences into various dependency subtrees of biological relations. We call all of these dependency subtrees "Evolutional Trees". Evolutional Trees are newly generated dependency paths that may include more sentence structures of the relation between two biological entities not covered by the training corpus. Consider the example in Figure 2. From these two dependency paths in the training corpus, new dependency paths can be learned by the proposed evolution operations of dependency trees. The constructed dependency tree structures are then used for the extraction of biological relations from literature and are also prospected to comprehensively represent the biological relations in literature from a small and limited training corpus.

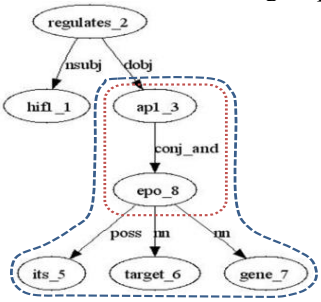


Figure 1. The evolution concept of dependency path from dependency parse tree.

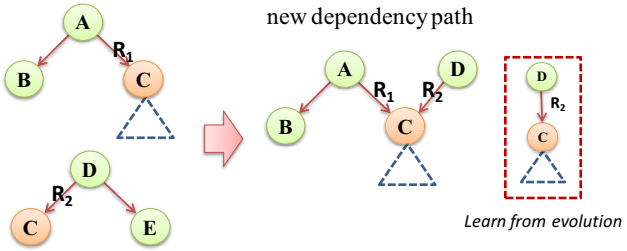


Figure 2. Learning from tree evolution

II. METHODS

A. System Overview

For the purpose of extracting biological relations from literature, we build an integrated framework as shown in Figure 3 to construct and utilize the evolutional trees from

the annotated corpus. In the pre-processing step, we applied the named entity recognition system AIIA-GMT [25] for identification all of the named entities in a sentence. We established a key verb set called Key Verb containing 190 verbs to indicate the biological relationships. Key Verb is expanded when another verb to describe the biological relationships is found.

Before the construction of evolutional trees, we must extract the dependency parse trees of named entities from the training corpus to form the seed tree set first. We use 100 positive sentences provided by biological experts as our original training set. All the positive sentences have gene regulatory relationships. We use Stanford Parser to parse these sentences into the dependency parse tree and then extract the dependency subtree instead of the dependency path. We suggest using a dependency subtree containing linguistic and structural information for determining biological relationships. A dependency subtree can be represented by lexical and syntactic subtrees, as shown in Figure 4. Therefore, all of the nodes and edges in the lexical subtree have corresponding nodes and edges in the syntactic subtree. To reduce the data sparseness, the node of the transcription factor is replaced with the symbol "TF" and the target gene is replaced with the symbol "TG".

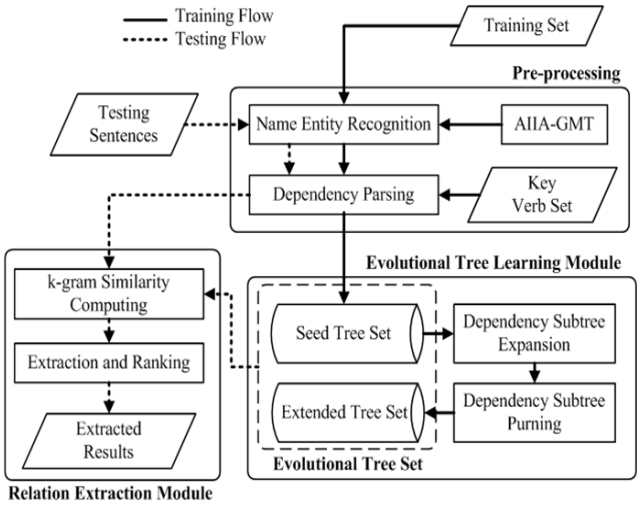


Figure 3. Overall framework of ETree

We assume that a verb and its voice are crucial for biological relationship extraction. Therefore, we apply our key verb set to identify verbs on the dependency subtree and classify the key verbs into verb and noun forms. The key verb with verb forms represents an active voice, such as "regulate", "induce" and "bind" and the key verb with noun forms represents a passive voice, such as "upregulation", "induction" and "interaction". If the key verb on the dependency subtree and the part of speech of the key verb is "NN" or "NNP", which means a noun compound modifier, it will be replaced with the symbol "N_Verb". On the contrary, if the key verb on the dependency subtree and the part of speech of the key verb is "VB", "VBD" or "VBZ", which means a verb modifier, it will be replaced with the symbol "V_Verb". An example is shown in Figure 4, where the key verb on the lexical subtree is "activation" and the part of speech of "activation" is "NN"; therefore, it is replaced with "N_Verb". Finally, we collect all the dependency subtrees that contain lexical subtrees and syntactic subtrees to form the Seed Tree Set.

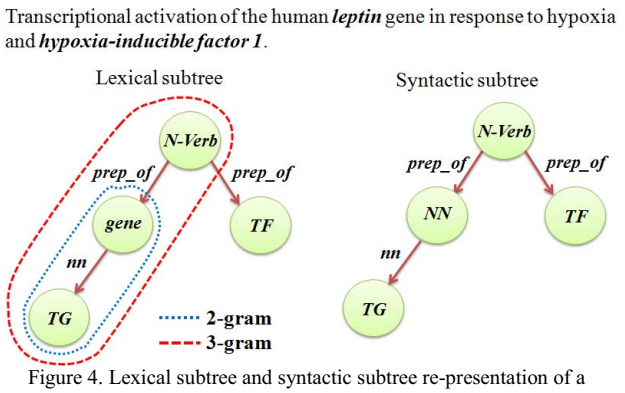


Figure 4. Lexical subtree and syntactic subtree re-presentation of a regulatory

B. *Evolutional Tree Learning*

The primary goal of the Evolutional Tree Learning Module is to generate the "Evolutional Tree Set". The Seed Tree Set consists of dependency subtrees in our original training set. Each dependency subtree has its corresponding lexical and syntactic subtrees. To formalize, let $T^L = (N^L, E^L)$ is a lexical subtree, and $T^S = (N^S, E^S)$ is a syntactic subtree, where N and E represent a set of nodes and edges, respectively. Assuming the number of subtrees in the training set is m, the Seed Tree Set $\boldsymbol{S}$ can be represented as $\{T^L_1, T^L_2, \ldots, T^L_m, T^S_1, T^S_2, \ldots, T^S_m\}$. For further operations on the tree set, we use a ***k-gram segment*** in a tree to be our operation unit. In a tree, a *k*-gram segment is a path with length of *k-1*. A *k*-gram segment contains *k* nodes and *k-1* edges and is represented as an ordered tuple $(n_i, e_{i,i+1}, n_{i+1}, \ldots, n_{i+k-2}, e_{i+k-2,i+k-1}, n_{i+k-1})$ where n_i and n_{i+k-1} are terminal nodes in the path. Consider the example tree in Figure 4, 2-gram segments in the lexical tree are (*gene*, *nn*, *TG*), (*N_verb*, *prep_of*, *gene*), and (*N_verb*, *prep_of*, *TF*). Two 3-gram segments are (*N_verb*, *pref_of*, *gene*, *nn*, *TG*) and (*gene*, *prep_of*, *N_verb*, *prep_of*, *TF*). Note that (*gene*, *nn*, *TG*) and (*TG*, *nn*, *gene*) are considered as the same segment in our segment operations.

Based on the tree set $\boldsymbol{S}$, we use two proposed tree operations, i.e., *expansion* and *pruning*, to construct the extended tree set that substantially generate and cover more dependency subtrees of biological relations. The proposed operations aim to find unknown and probable descriptions of biological relations that evolved from the original training set. In our observations, edges in dependency parse trees represent the type dependency relations between pairs of nodes and provide an effective and accurate description of the grammatical relationships in a sentence. There are several categories of the type dependencies, which are classified by Stanford Parser, and each category has its particular property. We adopt 10 categories defined in the

first level of categories of the type dependencies [11, 26] as our edge types. To represent the evolutional relations, we define that two different 2-gram segments are ***homologous*** if they have the same node value and the same edge value or similar edges with the same edge type. ***Homologous segments*** preserve the similar semantics if they appear in similar sentences. For example in Figure 5, the 2-gram segment (*dependent*, *nsubj*, *expression*) in (b) is homologous to segment (V_verb, *nsubjpass*, *expression*). They have the same node "*expression*" and their edges "*nsubj*" and "*nsubjpass*" belong to the same edge type. By the proposed expansion and pruning rules, we can generate new dependency subtrees from any two trees in ***S*** if they have one or more homologous segments.

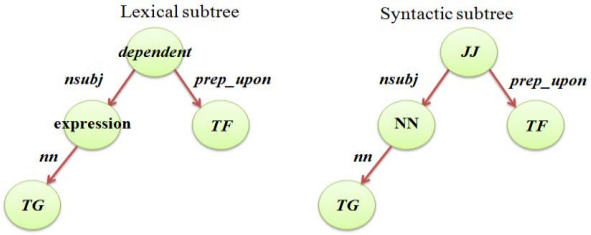
> ***Expansion*** -- *If the two dependency subtrees T_i and T_j in **S** have homologous segments, the system creates two different new dependency subtrees $T_{i,j}$ and $T_{j,i}$ from T_i and T_j. To construct $T_{i,j}$ ($T_{j,i}$), we join edge and another different node for each homologous segment in T_j (T_i) to the same node of the homologous segment in T_i (T_j).*

In general, $T_{i,j}$ is not equal to $T_{j,i}$. Consider the example of expansion in Figure 5, two example sentences are listed in Figure 5 (a) and Figure 5 (b). For the lexical subtree expansion, the lexical subtrees in Figure 5 have the homologous segments (dependent, prep_upon, TF) and (N_verb, prep_of, TF). Therefore, we joined the node "V_Verb" and the edge "prep_of" to the node "TF" of lexical subtrees of sentences (a) in Figure 5. Also, the edge "nsubj" of lexical subtrees of sentences (a) and the edge "nsubjpass" of lexical subtrees of sentences (b) in Figure 5 belong to the same edge type, which means a subject, and the same corresponding node "expression". Therefore, we joined the node "V_Verb" and the edge "nsubjpass" to the node "TF" of lexical subtrees of sentences (a) in Figure 5. Similarly, in the syntactic subtree expansion, we join the node "N_Verb" and the edge "prep_of" of the syntactic subtree of sentence (b) to the node "TF" of the syntactic subtree of sentence (a) in Figure 5. In addition, we joined the node "V_Verb" and the edge "nusbjpass" of the syntactic subtree of sentence (b) to the node "NN" of the syntactic subtree of sentence (a) in Figure 5. New lexical and syntactic subtrees are consequently generated as shown in Figure 6.

To understand the semantics of the new dependency subtree, we mapped it to the original sentence, as shown in Figure 6(c). The underlined words are nodes in the dependency subtree and the words in the brackets are the newly joined segments, which create a new sentence structure. The new subtree has a different meaning than that of the original dependency subtree. We can understand that the expression of the target gene is dependent on the transcription factor "HIF-1" and its upregulated expression relies on the activation of "HIF-1".

An example in Figure 7 shows the benefit of the expansion operation. The sentence in Figure 7(a) depicts that the expression of the target gene "VEGF" is upregulated by oxidative stressors by activation of the transcription factor "HIF-1". With the lack of similar training corpus, it is difficult to recognize the complicated relation description in this sentence. The lexical and syntactic subtrees in Figure 7 indicate some similar semantic relationships with the previous new dependency subtrees in Figure 6. As shown in Figure 7(b), the new dependency subtrees share semantic relationships of "TG expression", "up-regulated expression" and "activation of TF" of sentences in Figure 5(a) and Figure 5(b). Thus, by observing these semantic relationships, we may infer that the sentence structures of the two sentences are similar, and these extended semantic relations may help us to extract the regulatory relationship sentences not covered by the original dependency subtrees.

(a) Taken together, these results demonstrate that hypoxia induces ***leukocyte beta2 integrin*** expression and activity by transcriptional mechanisms dependent upon ***HIF-1***.



(b) Our results suggest that ***visfatin*** mRNA expression is upregulated in fat tissue through the activation of the ***HIF-1*** pathway due to hypoxia.

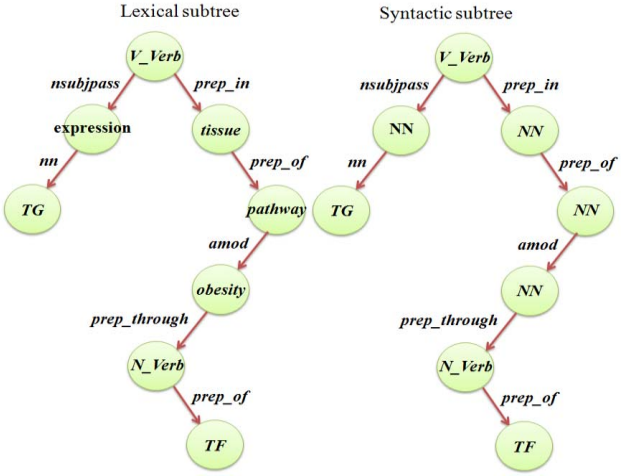
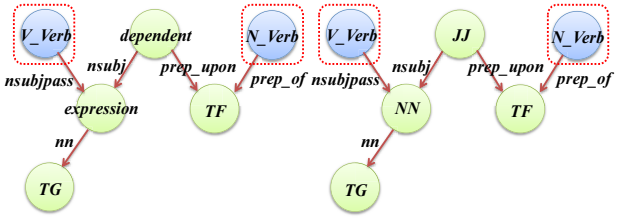


Figure 5. The example of two sentences for Dependency Subtree Expansion

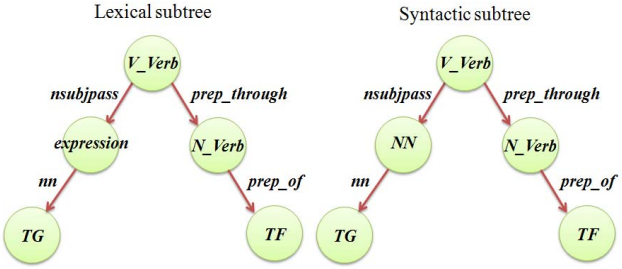
(a) The newly generated lexical subtree (b) The newly generated syntactic subtree



(c) Evolutional Sentence: "Taken together, these results demonstrate that hypoxia induces ***leukocyte beta2 integrin*** (*upregulated*) *expression* and functions by transcriptional mechanisms *dependent upon (activation of)* ***HIF-1***."

Figure 6. The new dependency subtrees and their semantic relationships

(a) Recent evidence suggests that ***VEGF*** expression is upregulated by oxidative stressors through activation of ***HIF-1***.



(b) TG expression, up-regulated expression, up-regulated through activation, activation of TF

Figure 7. The dependency subtree and the semantic relationships of the sentence

These new dependency subtrees may provide additional useful sentence structures, but some useless and illegal sentence structures will also be generated by the expansion. In these illegal structures, the new lexical subtree is different from the new syntactic subtree and does not follow the linguistic rules. We call it a "***Tree Inconsistency***" problem. To solve this problem, we propose a tree pruning method to automatically prune these unnecessary segments in the dependency subtrees. The pruning operation can make the lexical subtree consistent with the syntactic subtree and thus obey the linguistics rules.

Pruning -- *Intersect the newly generated lexical subtree* $T^{L}_{i,j}$ *and the newly generated syntactic subtree* $T^{S}_{i,j}$. *We make all of the nodes and edges in the lexical subtree have corresponding nodes and edges in the syntactic subtree. If any node and edge of* $T^{L}_{i,j}$ *are not in part of the intersection, we prune them in the dependency subtree. The pruned tree is denoted as* $T^{L'}_{i,j}$.

After the operations of expansion and pruning, the Extended Tree Set $\boldsymbol{X}$ can be represented as $\{T^{L}_{i,j}, T^{S}_{i,j} \mid \forall i,j\ (T^{L}_{i}, T^{L}_{j}) \vee (T^{S}_{i}, T^{S}_{j})$ *match the rules of Expansion*$\}$. Finally, the Seed Tree Set and Extended Tree Set are joined to form the Evolutional Tree Set $\boldsymbol{E}$ represented by $\boldsymbol{S} \cup \boldsymbol{X}$. The dependency subtrees in the Evolutional Tree Set are called "Evolutional Trees".

C. Dependency Subtree Weighting Strategy

We further propose a weighting strategy to assign a proper weight to each dependency subtree in the $\boldsymbol{X}$. First, we use each dependency subtree in $\boldsymbol{X}$ to extract all of the dependency subtrees in $\boldsymbol{S}$, except for its original dependency subtree in $\boldsymbol{S}$ because they have similar sentence structures. To calculate the proper weight of each dependency subtree, we do not take this case into account. The weight of dependency subtree T_{ij} in $\boldsymbol{X}$, denoted as $Weight_{ij}$, is shown below:

$$Weight_{ij} = \frac{\#\ positive\ sentences\ matched\ by\ T_{ij}\ \ from\ S}{|\ S - T_i\ |} \quad (1)$$

The weight represents the ratio that the dependency subtree can extract the positive sentences in $\boldsymbol{S}$. A higher weight means the larger number of sentences in $\boldsymbol{S}$ can be found by T_{ij}. This also indicates the dependency subtree T_{ij} conforms to the original training corpus and tends to be a better dependency subtree in Evolutional Tree Set. We use a threshold to cut off the dependency subtrees with lower weights. The filtering ratio is around 20% by adjusting the threshold in our weighting strategy. Finally, we use these remaining subtrees to extract biological relationships.

D. K-gram Similarity Measurement

We apply the Evolutional Tree Set and a k-gram similarity measurement to determine the similar sentence structure of the test sentence. In this similarity measurement, we calculate the ratio of the common grams between *k*-grams of two different trees for some *k*. If the longer segments are matched, we assume a higher probability of the sentence having a biological relationship. For example, t_1 is the dependency subtree in the Evolutional Tree Set, and t_2 is the dependency subtree of the test sentence. We use $G^{lex}_{k,i}$ to denote the set of all *k*-grams of the lexical subtree t_i^{lex} and $G^{syn}_{k,i}$ to denote the set of all *k*-grams of the syntactic subtree t_i^{syn}. We compute the ratio of the common *k*-grams of lexical subtrees as the lexical score $S^{lex}_{k,i}$ of t_1^{lex} and t_2^{lex}, as presented in Equation (2). The syntactic score $S^{syn}_{k,i}$ is calculated by the same way. In Equation (2), $C^{lex}_k(n_1, n_2)$ indicates the number of common k-grams of the two lexical subtrees, and $C^{syn}_k(n_1, n_2)$ is the number of common k-grams of the two syntactic subtrees. Then we use the average score $S_k(t_1, t_2)$ to evaluate the similarity between t_1 and t_2, as shown in following Equations.

$$S^{lex}_k(t^{lex}_1, t^{lex}_2) = \sum_{n_1 \in G^{lex}_{k,1}, n_2 \in G^{lex}_{k,2}} C^{lex}_k(n_1, n_2) \Big/ \left|G^{lex}_{k,1}\right|$$
$$S^{syn}_k(t^{syn}_1, t^{syn}_2) = \sum_{n_1 \in G^{syn}_{k,1}, n_2 \in G^{syn}_{k,2}} C^{syn}_k(n_1, n_2) \Big/ \left|G^{syn}_{k,1}\right| \quad (2)$$

$$S_k(t_1, t_2) = (S^{lex}_k(t^{lex}_1, t^{lex}_2) + S^{syn}_k(t^{syn}_1, t^{syn}_2))/2 \quad (3)$$

Finally, we use a threshold to predict whether the sentence has a biological relationship. If the similarity score of the sentence is greater than or equal to the threshold, the sentence is considered to have a biological relationship and is then extracted. If the similarity score is less than the threshold, we skip the sentence. However, there is an exception for sentence extraction. If the sentence is a compound-complex sentence, which has a semicolon, and each protein/gene of the protein/gene pair is in different simple sentences, we drop this sentence. The semicolon represents two relatively independent sentences in the compound-complex sentence and is used for internal division of the sentence. Therefore, no interactions or relationships are extracted when two proteins or genes are in different simple sentences.

III. RESULTS AND DISCUSSIONS

A. Data Sets

In our experiments, we focused on the extraction of two kinds of biological relationships, gene regulatory relationships and protein-protein interactions, to evaluate our proposed method. For gene regulatory relationships, we exploited the PubMed database to construct four datasets. Dataset D_A is built from manually selected articles by domain experts. The set of Dataset D_B (D_{B1}, D_{B2}, and D_{B3}) are built from search results of PubMed. We randomly chose three transcription factors, i.e., "AP-1", "E2F1" and "HIF-1", as our input queries to find related articles from PubMed. For each query, we chose the top 20 results and then selected the sentences related to these transcription factors. The correctness of the sentences was checked by biomedical experts. The statistics of the sentences are shown in TABLE I. For the protein-protein interaction data, we used the Learning Language in Logic (LLL) data set [11, 12]. The LLL data set is a publicly available data set that contains protein interaction annotations and has been used frequently in recent work on protein-protein interaction extraction. We used precision, recall and F-measures to evaluate the performance of our method.

TABLE I. The Statistics of Evaluation Datasets

Data Set	#sentences	#positive sentences	#negative sentences
Biological relation: gene regulation			
D_A	200	100	100
D_{B1}	270	100	170
D_{B2}	279	107	172
D_{B3}	619	241	378
Biological relation: protein-protein interaction			
LLL-train	269	134	135
LLL-test	61	30	31

B. Performance Evaluation

We first assess the performance of dependency subtrees and the expansion operation by using 100 positive sentences in D_A. There have 100 dependency subtrees (***S***) and 6938 extended trees (***X***) generated from these positive sentences. Total 200 sentences in D_A are used for testing. Note that the original tree of a sentence is removed from the similarity measurement when this sentence is evaluated. The similarity measurement is 2-gram measurement. The large amount of increasing on the recall rate in TABLE II shows that the representation of subtrees and the expansion of homologus segments effectively generate the useful information to extract the regulatory relations from a limited training corpus. In set *X*, there have 3746 trees that have the tree inconsistency problem. As shown in TABLE II, the pruning of these inconsistent segments will reduce the recall rate (-0.1) but highly increase the precision rate (+0.16).

TABLE II. Performance of Dependency Subtrees and Tree Operations

Method	Precision	Recall	F-measure
Evaluated on D_A			
Dependency Path	0.733 (11/15)	0.100	0.174
Dependency Subtree	0.730 (27/37)	0.247	0.369
+expansion	0.673 (68/101)	0.624	0.647
Evaluated on DB_{1-3}			
Dependency Subtree	0.612	0.465	0.529
+expansion	0.506	0.787	0.616
+pruning	0.660	0.685	0.672

In Figure 8, we evaluated several *k*-gram similarity measurements. The measurement *2-gram+3-gram* compare the union set of 2-gram segments and 3-gram segments and measurement *n-gram* compare the union set of all possible *k*-gram segments in a tree. It is reasonable that longer segments of a tree preserve more precise information of this tree and shorter segments generalize the semantics of this tree. The higher precision and lower recall rates of 3-gram measurement conform to this property with the lower precision and higher recall rates of 2-gram measurements. In general, the depth of a dependency subtree is about 3. A 3-gram segment is the most applicable segment to precisely represent the subtree segment. According to the occurrence frequency of segment, we listed top-6 3-gram segments in the evolutional tree set of gene regulation corpus in Figure 9. The occurrence frequencies of the 4-th and 6-th 3-gram segments are not so high in the original seed set. Many of these two kinds of segments are generated after the expansion operation. In experimental observation, these automatically generated segments are useful for the relation recognition of gene regulation.

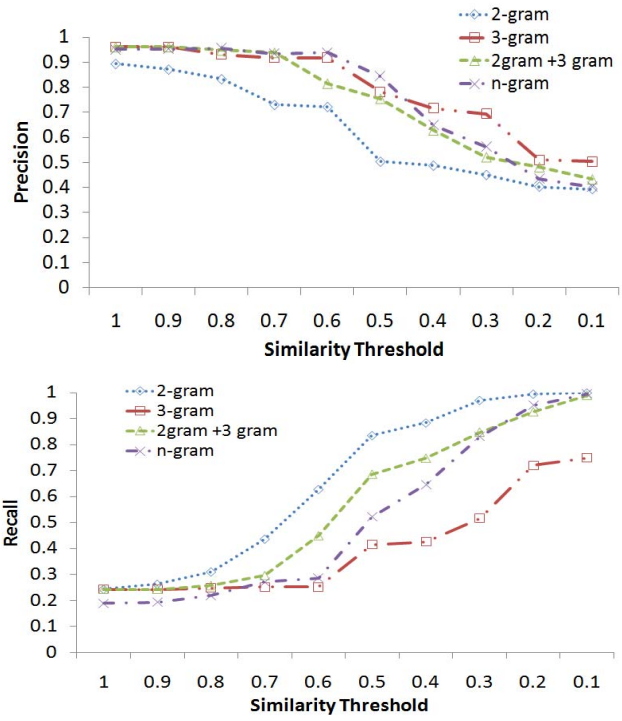


Figure 8. Performance comparison of *k*-gram similarity measurement

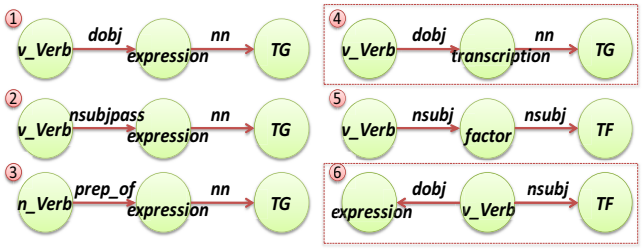


Figure 9. Top-6 3-gram segments in the evolutional tree set of gene regulation corpus

C. *Comparison with Other Systems*

According to our preliminary evaluation, the cut threshold of ETree was set to 0.5 for all experiments in this paper. In the preliminary evaluation, we use the combination of *2-gram and 3-gram* similarity measurements and set the similarity threshold 0.5 to filter the highly dissimilar gene regulation sentences. Three PubMed data sets were used to evaluate ETree and implemented three relation extraction methods. We used a rule-based method, which applies Stanford Parser to extract biological relations from biological text documents [9]. We also used the RelEx system [10] and the four kernel methods [23] for the performance comparison. The four kernal methods are the lexical kernel, POS kernel, syntactic kernel and Full kernel, which use lexical, POS and syntactic information on the dependency path between two entities. In addition, we also compared ETree with previous related work called "AutoPat" [27], which is a pattern-based approach to the extraction of gene regulatory relationships. The experimental results of the three PubMed data sets are shown in TABLE III.

TABLE III. Performance Comparison with Other Systems on Three PubMed Data Sets

Data	Method	Precision	Recall	F-measure
D_{B1}	Rule-Based	0.64	0.43	0.52
	Autopat	0.38	0.71	0.49
	RelEx	0.51	0.49	0.50
	K_Lex	0.60	0.58	0.59
	K_POS	0.51	0.76	0.61
	K_Syn	0.49	0.69	0.58
	K_Full	0.59	0.68	0.63
	ETree	**0.74**	0.67	**0.71**
D_{B2}	Rule-Based	0.66	0.40	0.51
	Autopat	0.41	0.85	0.55
	RelEx	0.53	0.53	0.53
	K_Lex	0.63	0.63	0.63
	K_POS	0.47	0.75	0.58
	K_Syn	0.49	0.80	0.61
	K_Full	0.53	0.74	0.62
	ETree	**0.77**	0.70	**0.73**
D_{B3}	Rule-Based	0.54	0.48	0.50
	Autopat	0.40	0.83	0.54
	RelEx	0.49	0.56	0.52
	K_Lex	0.62	0.62	0.62
	K_POS	0.51	0.62	0.56
	K_Syn	0.50	0.63	0.56
	K_Full	0.56	0.62	0.59
	ETree	**0.75**	0.73	**0.74**

ETree achieved the best performance on the three PubMed data sets, showing that our proposed evolutional dependency parse trees help to extract a variety of sentence structures. The rule-based method exhibited the worst performance because their predefined seven rules from the dependency relations, which are parsed by the Stanford Parser, are not appropriate for our data sets. The AutoPat system had the highest recall rate because they only consider two entities and the verbs that appear in the sentence. However, not all of the sentences with the above three elements really described a gene regulatory relationship. The RelEx system had a lower performance on our data sets likely because their method focuses on protein-protein interaction extraction, and their applied sets of rules are not appropriate for our data sets. Next, we compared ETree with other recent relation extraction systems that were trained and tested on the LLL Data Set. We used the results of previous works [8, 10, 22]. The experimental results of the LLL data set are shown in TABLE IV.

ETree achieved the highest precision rate and the best performance among all of the relation extraction systems. Therefore, our proposed evolutional dependency parse trees can deal with different types of relations and still predict the interactions between two proteins with high accuracy. However, the recall rates of our system are not the best. Because some positive and negative sentences have similar sentence structures, we raised the threshold value to get higher accuracy when predicting if a sentence has a biological relationship.

TABLE IV. Performance Comparison with Other Systems on The LLL Data Sets

Method	Precision	Recall	F-measure
Autopat	0.58	0.63	0.60
Rule-based	0.56	0.30	0.39
RelEx	0.82	0.72	0.77
Graph kernel	0.73	0.87	0.77
Predicate Kernel	0.71	0.72	0.72
Walk Kernel	0.73	0.83	0.78
Dependency Kernel	0.59	0.63	0.61
Hybrid Kernel	0.66	0.74	0.70
ETree	**0.88**	0.73	**0.80**

IV. CONCLUSIONS

In this work, we developed a biological relation extraction system, ETree, which is not restricted to particular types of biological relationships. We focused on the extraction of gene regulatory relationships and protein-protein interaction relationships. We used a dependency subtree rather than a dependency path on the dependency parse tree for learning the tree information of the biological relationships. Due to the limited training corpus, we propose a novel method called "Evolutional Tree" to generate different sentence structures. Our method achieved promising results on three PubMed data sets and the LLL data set. Moreover, ETree can be applied to different kinds of relation extractions, e.g., gene regulatory relationships and protein-protein interactions.

REFERENCES

[1] A. Ozgur, *et al.*, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics,* vol. 24, pp. i277-85, Jul 1 2008.

[2] J. Hankenberg, *et al.*, "Molecular event extraction from link grammer parse trees," in *The Workshop on BioNLP*, 2009.

[3] E. Buyko, *et al.*, "Testing Different {ACE}-Style Feature Sets for the Extraction of Gene Regulation Relations from {MEDLINE} Abstracts," in *The Third International Symposium on Semantic Mining in Biomedicine (SMBM)*, Turku, Finland, 2008, pp. 21-28.

[4] J. Saric, *et al.*, "Large-scale extraction of gene regulation for model organisms in an ontological context," *In Silico Biol,* vol. 5, pp. 21-32, 2005.

[5] J. Saric, *et al.*, "Extraction of regulatory gene/protein networks from Medline," *Bioinformatics,* vol. 22, pp. 645-50, Mar 15 2006.

[6] G. Erkan, *et al.*, "Semi-Supervised Classification for Extracting Protein Interaction Sentences using Dependency Parsing," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 228-237.

[7] T. Fayruzov, *et al.*, "DEEPER: A Full Parsing Based Approach to Protein Relation Extraction," presented at the Proceedings of the 6th European Conference, EvoBIO, 2008.

[8] A. Airola, *et al.*, "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning," *BMC Bioinformatics,* vol. 9 Suppl 11, p. S2, 2008.

[9] D. Lipika and R. Goyal, "Using Relations to Index Biological Document Repositories for Efficient Searching," presented at the International Conference on Management of Data COMAD, 2006.

[10] K. Fundel, *et al.*, "RelEx--relation extraction using dependency parse trees," *Bioinformatics,* vol. 23, pp. 365-71, Feb 1 2007.

[11] S. Pyysalo, *et al.*, "Comparative analysis of five protein-protein interaction corpora," *BMC Bioinformatics,* vol. 9 Suppl 3, p. S6, 2008.

[12] C. Nedellec, "Learning language in logic - genic interaction extraction challenge.," in *In Proceedings of the ICML05 workshop: Learning Language in Logic (LLL05)*, 2005.

[13] J. Bjorne, *et al.*, "Extracting Complex Biological Events with Rich Graph-Based Feature Sets," in *Preceedings of the Association for Computational Linguistics (ACL)*, Boulder, Colorado, 2009, pp. 10-18.

[14] J. D. Kim, *et al.*, "GENIA corpus--semantically annotated corpus for bio-textmining," *Bioinformatics,* vol. 19 Suppl 1, pp. i180-2, 2003.

[15] C. B. Giles and J. D. Wren, "Large-scale directional relationship extraction and resolution," *BMC Bioinformatics,* vol. 9 Suppl 9, p. S11, 2008.

[16] Y. Miyao, *et al.*, "Evaluating contributions of natural language parsers to protein-protein interaction extraction," *Bioinformatics,* vol. 25, pp. 394-400, Feb 1 2009.

[17] G. Zhou, *et al.*, "Tree Kernel-Based Relation Extraction with Context-Sensitive Structured Parse Tree Information," in *the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 728-736.

[18] F. Reichartz, *et al.*, "Dependency Tree Kernels for Relation Extraction from Natural Language Text," presented at the Proceedings of the ECML PKDD, 2009.

[19] V. Truc, *et al.*, "Convolution Kernels on Constituent, Dependency and Sequential Structures for Relation Extractioin," in *The Conference on Empirical Methods in Natural Language Processing and Association for Computational Linguistics*, 2009, pp. 1378-1387.

[20] J. Li, *et al.*, "Kernel-based learning for biomedical relation extraction," *Journal of the American Society for Information Science and Technology,* vol. 59, pp. 756-769, 2008.

[21] Z. Min, *et al.*, "Exploring syntactic structured features over parse trees for relation extraction using kernel methods," *Information Processing & Management,* vol. 44, pp. 687-701, 2008.

[22] S. Kim, *et al.*, "Kernel approaches for genic interaction extraction," *Bioinformatics,* vol. 24, pp. 118-26, Jan 1 2008.

[23] T. Fayruzov, *et al.*, "Linguistic feature analysis for protein interaction extraction," *BMC Bioinformatics,* vol. 10, p. 374, 2009.

[24] E. Buyko, *et al.*, "Event extraction from trimmed dependency graphs," in *BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, 2009.

[25] C. N. Hsu, *et al.*, "Integrating high dimensional bi-directional parsing models for gene mention tagging," *Bioinformatics,* vol. 24, pp. i286-94, Jul 1 2008.

[26] C. Manning, "Generating typed dependency parses from phrase structure parses," presented at the 5th International Conference on Language Resources and Evaluation, 2008.

[27] Y. T. Tang, *et al.*, "Using Unsupervised Patterns to Extract Gene Regulation Relationships for Network Construction," *PLoS One,* vol. 6, p. e19633, 2011.