Wuhan University
Journal of Natural Sciences

2015, Vol.20 No.2, 141-145

Article ID 1007-1202(2015)02-0141-05
DOI 10.1007/s11859-015-1072-2

Semantic Relation Annotation for Biomedical Text Mining Based on Recursive Directed Graph

☐ CHEN Bo [1,2], LÜ Chen[1], WEI Xiaomei [1], JI Donghong [1†]

1. School of Computer, Wuhan University, Wuhan 430072, Hubei, China;

2. Department of Chinese Language and Literature, Hubei University of Art and Science, Xiangyang 441053, Hubei, China

Abstract: In this paper we propose a novel model "recursive directed graph" based on feature structure, and apply it to represent the semantic relations of postpositive attributive structures in biomedical texts. The usages of postpositive attributive are complex and variable, especially three categories: present participle phrase, past participle phrase, and preposition phrase as postpositive attributive, which always bring the difficulties of automatic parsing. We summarize these categories and annotate the semantic information. Compared with dependency structure, feature structure, being recursive directed graph, enhances semantic information extraction in biomedical field. The annotation results show that recursive directed graph is more suitable to extract complex semantic relations for biomedical text mining.

Key words: biomedical text mining; semantic annotation; recursive directed graph; postpositive attribute

CLC number: TP 301, H 085

Received date: 2014-10-15

Foundation item: Supported by the National Natural Science Foundation of China (61202193, 61202304), the Major Projects of Chinese National Social Science Foundation (11&ZD189) and the Chinese Postdoctoral Science Foundation (2013M540593, 2014T70722)

Biography: CHEN Bo, female, Ph.D., Associate professor, research direction: natural language processing. E-mail: chenbo@ whu.edu.cn

† To whom correspondence should be addressed. E-mail: dhji@ whu.edu.cn

0 Introduction

Annotating biomedical text plays an important role in the fields of biomedical text mining and information extraction. It is increasing the accuracy and efficiency in automatic retrieval [1-3].

However, the resources without semantic information bring the problems to recognize the entity and extract the key words, which are what the doctors need urgently, such as Gene Epigenetics, Oncology.

In recent years, the semantic annotation becomes more important in biomedical annotation field [4-6]. In this paper, we propose a new method "recursive directed graph", which is a semantic representation model for biomedical text mining. The method is satisfactory to represent or derive the biomedical conceptual relations of biomedical complex sentence patterns. We focus on building a large scale labeled biomedical resource, the biomedical token semantic association (bioTSA), consisting of part of BioNLP2009 ST and BioNLP2013 GE ST training, which can describe all the semantic relations of tokens in the text.

Currently, dependency structure is one of the most popular representation methods. Many researches for parsing text have been done successfully in this way [7, 8]. Other relative annotation researches, such as the framework developed by Kulick *et al* [2], which integrates the Treebank and the Propbank, contains syntactic structure and predicate-argument structure; the new concepts "Single-facet Annotation and Semantic Typing" provided by Kim *et al* [3] for semantic annotation and event annotation.

However, many problems are encountered in parsing biomedical text, in which there are many special sentence patterns, such as postpositive attributive, inverted sentences, the complex noun phrase, the verb-complement structure, etc. It is difficult to find the correct head, which leads to errors extracting entity relations.

We put forward a new method "recursive directed graph" for parsing biomedical text. In previous work, we already built a large-scale semantic resource with 30 000 Chinese sentences with feature structure in three years. It enriches Chinese semantics resources [9]. It is an attempt to use "recursive directed graph" in annotation of English biomedical text.

In this paper, we choose postpositive attributive as the object. Section 1 discusses the method we propose. Section 2 focuses on the annotation of postpositive attributive sentences in biomedical text. Section 3 is the discussion, includes the overall annotation research, including labeled data, annotators, and the consistency of annotation. Section 4 is the conclusion.

1 Annotation with Recursive Directed Graph

Feature structure is not a new term, which is common in many fields, such as generative phonology [10], generalized phrase structure grammar (GPSG) [11], lexical functional grammar (LFG) [12]. We borrow the term "Feature Structure" to provide a new model, which can be formalized "recursive directed graph", allowing a more complete semantic description for biomedical text [9]. We focus on the better representation of semantic relations. Recursive directed graph can be shown in Fig.1.

Generally, a phrase or sentence may be expressed as a collection of feature structures, and a feature structure is represented as a triple: [Entity, Feature, Value]

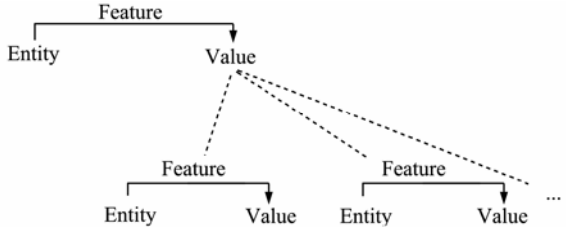

Fig. 1 Feature structure: recursive directed graph

A triple can be regarded as two nodes and the edge that links them. A node is an entity or a value, and the edge is a feature. The feature must be the feature of one of the nodes, and the node serves as the owner of the feature, while another node serves as the value. Thus, a feature structure can be seen as a directed graph. Because the value can also be another feature structure, the feature structure may be represented as a recursive graph, in which a node can also be a graph [13].

Feature structure, as "recursive directed graph", allows multiple semantic links and multiple nesting. According to previous researches [9,13], it is more suitable for extracting complex semantic relations.

Example 1 *gene* ***expression*** *from the HTLV-I LTR* (1)

Example (1) is complex noun phrase with a preposition phrase. It is very common in biomedical text. The entity is "gene expression", the feature is "from", and the value is "the HTLV-I LTR". (1) can be described by three triples, Figure 2 is the feature structure graph of (1).

Triple1-1: [expression, , gene];

Triple1-2: [expression, from, the HTLV-I LTR];

Triple1-3: [HTLV-I LTR, , the].

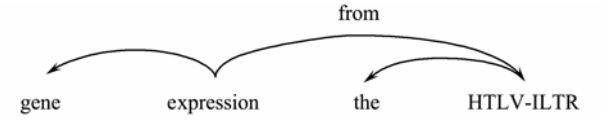

Fig. 2 The feature structure graph of example (1)

Example 2 ***Regulation*** *of T*1 *expression during induction of monocytic differentiation by okadaic acid* (2)

Example (2) is the title of a paper, which is a complex noun phrase with serial nouns. The sentence structure is more complex than (1), in which the semantic relations are interrelated and complex. (2) can be described by 6 triples:

Triple2-1: [regulation, during, induction];

Triple2-2: [regulation, of, expression];

Triple2-3: [induction, of, differentiation];

Triple2-4: [differentiation, by, okadaic acid];

Triple2-5: [expression, , T1];

Triple2-6: [differentiation, , monocytic].

In triple2-2, "expression" is the value of the entity "regulation", meanwhile, in triple2-5, "expression" is the entity, whose value is "T1". And "differentiation" has the same situation. Therefore, in feature structure model, one node can be multiple-semantic relations node. Figure 3 is the feature structure graph of (2).

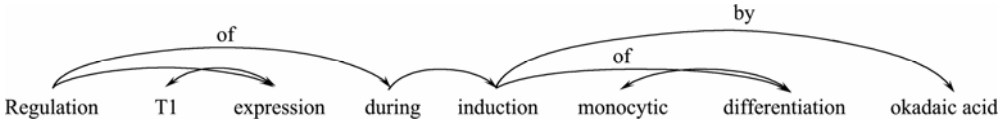

Fig. 3 The feature structure graph of (2)

2 Semantic Annotation of Postpositive Attributive Sentence Patterns in Biomedical Text

Like the adjectives, the function of postpositive attributive aims to modify and describe nouns or noun phrases [14]. The usages of postpositive attributive are complex and variable.

Postpositive attributive sentence pattern in biomedical text is very common. In syntax, there are three types: First, the clause as postpositive attributive, as in *who*, *whom*, *which*, *whose*, etc.; Second, the phrase as postpositive attributive, as in infinitive phrase, present participle phrase, past participle phrase, adjective phrase, preposition phrase, etc.; Third, a single word as postpositive attributive.

In all the types, present participle phrase, past participle phrase, and preposition phrase as postpositive attributive always bring the difficulties of automatic parsing. It is easy to get confused to find the correct head noun which the postpositive attributive modifies, then lead to errors extracting entity relations. We have annotated 113 biomedical documents and 906 sentences, in which there are 82 postpositive attributive sentences, the proportion is 9%. (3)-(5) are the typical examples in data:

Example 3 *T*10 *mRNA levels were superinduced in cells treated with both okadaic acid and cycloheximide, whereas inhibition of protein synthesis had little*, *if any, effect on okadaic acid-induced T*11 *transcription*. (3)

In (3), the postpositive attributive is the past participle phrase "treated with both okadaic acid and cycloheximide", the head noun is "cell". The semantic relation is "patient-predicate". The postpositive attributive in (3) can be described by three triples, Fig. 4 is the feature structure graph of (3).

Triple3-1: [treated, ,cells];

Triple3-2: [treated, with, okadaic acid];

Triple3-3: [treated, with, cycloheximide].

Example 4 *Suppression of signals required for* ***activation*** *of* ***transcription factor*** *NF-kappa B in* ***cells constitutively expressing*** *the HTLV-I Tax protein* (4)

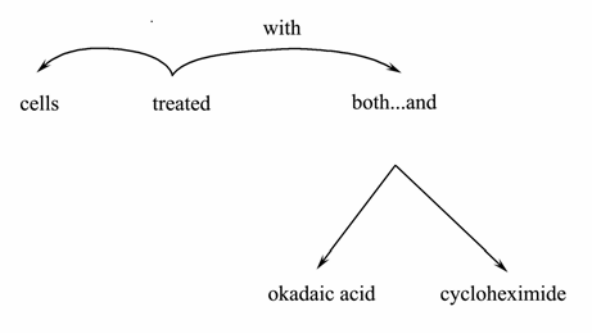

Fig. 4 The feature structure graph of (3)

In (4), the postpositive attributive is the present participle phrase "constitutively expressing the HTLV-I Tax protein". But the head noun is uncertain, the three nouns can be the head: "activation", "transcription factor", and "cells". If just considering the distance, it maybe "cells". However, the head should be the "transcription factor". The postpositive attributive in (4) can be described by six triples, Fig. 5 is the feature structure graph of (4).

Triple4-1: [expressing, , the HTLV-I Tax protein];

Triple4-2: [expressing, , constitutively];

Triple4-3: [expressing, , transcription factor];

Triple4-4: [transcription factor, , NF-kappa B];

Triple4-5: [transcription factor, in, cells];

Triple4-6: [activation, of, transcription factor].

Example 5 *In contrast*, *in a number of multiple myeloma cell lines*, *representing differentiated*, *plasma cell-like B cells*, ***PU*.1 *DNA binding activity***, ***mRNA expression, and Pu box-dependent transactivation*** *were absent or detectable at a very low level* (5)

In (5), it is hard to ensure the objects of the postpositive attributive verb "binding". It is just "activity", or "activity, mRNA expression", or "activity, mRNA expression, and Pu box-dependent transactivation". According to the semantic annotation, the subject of "binding" is "DNA", its object should be "activity". The postpositive attributive in (5) can be described by three triples, Figure 6 is the feature structure graph of (5).

Triple5-1: [DNA, , PU.1];

Triple5-2: [binding, , DNA];

Triple5-3: [binding, , activity].

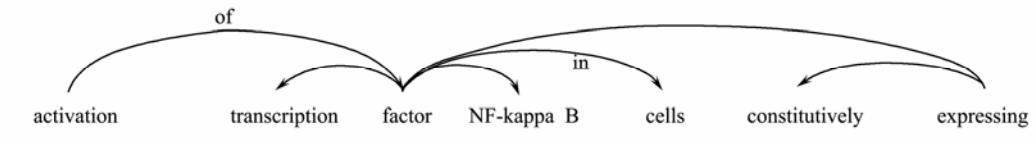

Fig. 5 The feature structure graph of (4)

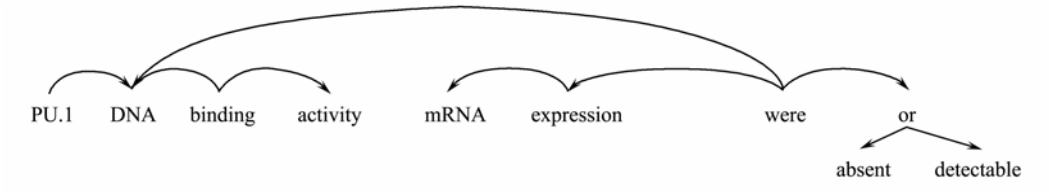

Fig. 6 The feature structure graph of (5)

Postpositive attributive is more error-prone than other sentence patterns. We just annotated 82 postpositive attributive sentences, and summarized the main three types. Using feature structure model can resolve these problems, and can represent more semantic information from biomedical texts than traditional dependent structure.

3 Discussion

We selected 113 text materials, 11 abstracts from BioNLP'09 ST, and 102 documents from BioNLP2013 GE task. We construct a small biomedical semantic resource with 906 sentences, and focus on annotating semantic relations of sentences.

The group of annotators includes 20 masters and doctors with linguistics, bioinformatics and computer engineer backgrounds at Wuhan University and Huazhong Agricultural University. Annotation training consisted of the annotation method, the annotation standard, the annotation platform, and annotated examples. The annotators worked independently, the consistency of annotation achieves 95%, which is very good because feature structure only requires determining the semantic relations. We would crosscheck the results periodically to avoid manual errors every week.

We compare feature structure with Stanford parser [15] to parse the five example sentences. Table 1 shows the results of annotation precision and recall, based on the online Stanford parser and feature structure. The online Stanford parser gets three correct and two incorrect results. When the parser encounters the more complex postpositive attributive structures, it is hard to ensure the right semantic pairs.

Table 1 Result of the online Stanford parser in comparison with feature structure

Example	Correct annotation of the postpositive attributive	Annotation result of online stanford parser	Result
Example 1	[expression, from, the HTLV-I LTR]	prep_from(expression-2, LTR-6)	Correct
Example 2	[differentiation, by, okadaic acid]	prep_by(Regulation-1, acid-12)	Wrong
Example 3	[treated, ,cells]	vmod(cells-7, treated-8)	Correct
Example 4	[expressing, , transcription factor]	nsubj(expressing-15, activation-6)	Wrong
Example 5	[binding, , activity]	amod(activity-25, binding-24)	Correct

4 Conclusion

The novel model "feature structure" that we put forward is formalized "recursive directed graph" for the semantic representation. It is a successful attempt to use the method in biomedical text. In future work, we will expand the biomedical corpus. Compared with other models, feature structure is more suitable for extracting biomedical complex semantic relations, and can represent more semantic relations and allows multiple links. According to the results, labeling with feature structures is much more expeditious and effective than dependency structures. In the application, our research is significant to biomedical text mining by providing rich semantic information. The resource can be used directly to relation extraction, event extraction, and automatic question and answering.

References

[1] Pyysalo S, Ginter F, Heimonen J, *et al*. BioInfer: A corpus for information extraction in the biomedical domain [J]. *BMC Bioinformatics*, 2007, **8**(1): 50.

[2] Kulick S, Bies A, Liberman M, *et al*. Integrated annotation for biomedical information extraction[C]//*Proc of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics* (*HLT/NAACL*). Boston: Association for Computational Linguistics, 2004: 61-68.

[3] Kim J D, Ohta T, Tsujii J. Corpus annotation for mining

biomedical events from literature [J]. *BMC Bioinformatics*, 2008, **9**(1): 10.

[4] Akane Y S, Yusuke M Y, Yuka Tateisi Y K, *et al*. Biomedical information extraction with predicate-argument structure patterns[C]//*Proceedings of the First International Symposium on Semantic Mining in Biomedicine* (*SMBM*). Budapest: CEUR, 2005.

[5] Spasic I, Ananiadou S, McNaught J, *et al*. Text mining and ontologies in biomedicine: Making sense of raw text [J]. *Briefings in Bioinformatics*, 2005, **6**(3): 239-251.

[6] Cohen A M, Hersh W R. A survey of current work in biomedical text mining[J]. *Briefings in Bioinformatics*, 2005, **6**(1): 57-71.

[7] Zhang Y, Nivre J. Transition-based dependency parsing with rich non-local features[C]//*Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*: *Short Papers-Volume* 2. Boston: Association for Computational Linguistics, 2011: 188-193.

[8] Mel′čuk I. *Dependency Syntax*: *Theory and Practice* [M]. Herndon: SUNY Press, 1988.

[9] Chen B, Wu H M, Lv C, *et al*. Semantic labeling of Chinese serial verb sentences based on feature structure [J]. *Lecture Notes in Computer Science*, 2013, **8229**(1): 784-790.

[10] Kenstowicz M, Kisseberth C. *Generative Phonology* [M]. New York: Academic Press, 1979.

[11] Gazdar G. *Generalized Phrase Structure Grammar* [M]. Cambridge: Harvard University Press, 1985.

[12] Dalrymple M. *Lexical Functional Grammar* [M]. New York: Academic Press, 2001.

[13] Chen B, Ji D, Lv C. Building a Chinese semantic resource based on feature structure [J]. *International Journal of Computer Processing of Languages*, 2012, **24**(1): 95-101.

[14] Lu J, Lu K. Research on syntactic characteristics of computer English and its English to Chinese translation strategy[C]//*Proc of* 2013 *Fifth International Conference on the Computational and Information Sciences* (*ICCIS*). Los Alamitos: IEEE Computer Society, 2013: 1867-1870.

[15] de Marneffe M C, MacCartney B, Manning C D. Ctenerating typed dependency parses from phrase structure parses[C]//*Proceedings of LREC*. Paris: EIRA, 2006: 449-454.

□