The Application of Dipre in the Calculation of Coherence Relations

Wei Jiang

Communication University of China, Beijing, China

Keywords: Dipre, Computability, Coherence Relations

Abstract: at present, natural language processing has made some achievements in terms of vocabulary, syntax and semantic processing; some mature theoretical models and analysis methods have been established. However, the research of discourse is still in its infancy, which greatly restricts the further development and promotion of natural language processing. This paper attempts to analyze the computability of coherence relations through the dipre algorithm, in order to explore and prospect the research approach of discourse coherence calculation.

1. Introduction

Recently, natural language processing has made great progress and achievements. But it still faces many problems like difficulties in knowledge integration. Currently, the knowledge construction, the knowledge representation and the knowledge reasoning are still in a relatively primary stage; problems that can be solved are limited. Human beings have strong representation abilities for encyclopedia knowledge, so it is difficult for computers to establish a knowledge base which covers all human cognition. In addition, the discourse world implies certain information, including the time, the place, the environment and other factors. At present, the context processing and reasoning are relatively elementary compared with human intelligence; words that are not difficult for human beings to understand can not be understood by computers. For example, one person said “hurry up”; his partner said “it's only 10”. This scene is easy to understand, if we suppose that this scene happens at 8:10, and the time agreed by the two persons to arrive at a certain place is 8:20. But the computer can't comprehend that. In other words, human beings can understand the environmental background well and acquire and use knowledge flexibly in the process of communication, but it is difficult for computers to do so.

At present, there are two kinds of typical discourse relational corpuses: the Rhetorical Structure Theory Discourse Treebank based on the RST theory and the Penn Discourse Tree Bank based on the PDTB system. They adopt different relational type systems and annotation standards. The existing theories on the corpus and annotation mainly focus on English. Domestic scholars have done some analyses on Chinese study. However, due to the characteristics of Chinese and the differences between Chinese and English in semantics, semantic relations, tenses and other aspects, we can not directly translate the English relational type systems into Chinese. Therefore, it is urgent to build a Chinese oriented discourse coherence type system so as to better serve the study of Chinese.

The DIPRE (Dual Iterative Pattern Relation Extraction) is an algorithm proposed by Sergey Brin, one of the founders of Google, for extracting data of specific formats or types on the Internet. It can be used for relation extraction in natural language processing. The success of Google search engine pushes the processing of natural language to a new height. Different hierarchies focus on different points and deal with different ways. Based on the task of building a corpus of coherence relations, this paper uses the DIPRE algorithm to calculate the coherence relations, so as to categorize the coherence relations and finally realize the automatic recognition of coherence relations in Chinese texts.

2. The Computability of Coherence Relations

Computational linguistics refers to the process of “analyzing and processing natural languages

DOI: 10.25236/iiicec.2019.039

by establishing formal mathematical models, and realizing the analysis and processing through programs in the computer, so as to achieve the purpose of simulating all or a part of human language ability by machines”. That is to say, computability means to solve a certain kind of practical problems by the computer; a computable problem means it can be formally represented by computers in limited steps and solved by computers. Specifically, the computability at the discourse level mainly deals with the processing and analyzes of the meaning of natural discourses on the basis of surface form features of the language.

Since the 1960s and 1970s, the study of discourse coherence has developed into several influential and highly accepted theoretical models at home and abroad. Examples include Van Dijk's Macro Structure Theory, Halliday and Hasan's cohesion theory proposed in *Cohesion in English*, Hobbs's Coherence Theory, as well as Mann and Thompson's Rhetorical Structure Theory (RST) proposed in *Rhetorical Structure Theory: A Theory of Text Organization*. They have different understanding and handle the coherence relations in different ways; the issue is explained from different levels. In the final analysis, in the context of computational linguistics, the computability of coherence depends on the computer recognition and processing of discourse coherence.

3. The Principle of Dipre Algorithm

The main steps of DIPRE algorithm are as follows. The first is to extract patterns from known seeds; the second is to extract information from unknown texts through the obtained patterns; the third is to extract new patterns from the extracted information; the fourth is to apply them into new texts. These steps move in circles, so that the “snowballing” information amplification can be realized. It is a representative method in semi supervised relation extraction.

After determining the relationship to be extracted, the first task is to provide seeds. For example, to investigate the cultural market, we hope to comprehensively analyze factors influencing the rising box office of domestic cinemas, and to extract the relations between the rising box office and corresponding factors based on relevant information in web news. After the research objectives are determined, a set of seeds are given for the relationship to be studied: (Cinema facilities are constantly improved, box office). The relationship between the continuous improvement of facilities and the satisfaction of box office: the box office of cities below the third-tier has increased significantly due to the continuous improvement of theater facilities in different regions.

After obtaining the seed, we need to determine the tuple information. This step is crucial. It determines what information is extracted based on the text and whether a fixed pattern can be extracted. If we can't extract the fixed pattern, we can't extract the new relationship, that means the self increasing of information can't continue. To be specific, determining the tuple information in advance can make the processing proceed in a fixed mode, so as to reduce the complexity of calculation; without specifying the tuple information in advance, other methods can also be used to mine the pattern of text information containing seeds to obtain the relationship pattern of seed information. Both methods have advantages and disadvantages. The first method requires the intervention of expert knowledge to define the relationship characteristics of relevant elements. The second method does not need the intervention of expert knowledge, but it has higher complexity and lower expected value. In this paper, the first method is used to determine the tuple manually, and then extract the corresponding tuple information from the text.

After determining the tuple information, it is necessary to extract the tuple relation from statements containing seeds in the input text. In the example, we suppose to determine a six tuple information [order, entity 1, entity 2, prefix, suffix, middle]. In the tuple, order represents the relative ordinal relation between entity1 and entity 2 in the statement. If entity 1 is in front, the order will be 1; otherwise it is 0. Entity 1 refers to the first entity in the relationship seed, which means the “continuous improvement of cinema facilities”. Entity 2 represents the second entity in the relationship seed, which is the “box office”. The prefix represents the morphemes immediately preceding the entity 1 and entity 2 constituent pairs. The suffix represents the morpheme immediately following the entity 1 and entity 2 pairs. The middle represents the morpheme connecting entity 1 and entity 2.

The specific application of the DIPRE algorithm is illustrated by the following statement. Input sentence 1: because of the continuous improvement of theater facilities in various regions, correspondingly box office of cities below the third-tier has increased significantly. The results of information extraction are as follows.

6 tuple: [order, entity 1, entity 2, prefix, suffix, middle] = [1, the continuous improvement of theater facilities in various regions, box office, because of, increased significantly, correspondingly]. In sentence 1, the continuous improvement of theater facilities appears in front of the box office, so the order is 1; “because of” appears in front of entity 1, “the continuous improvement of theater facilities in each region” and entity 2, “the box office”, so it is a prefix. Similarly, “increased significantly” is the suffix. The conjunction between two entities is “correspondingly”, which corresponds to the middle. Input sentence 2: In recent years, the gradual improvement of theater facilities in various regions has led to a significant increase in box office in cities below the third-tier. The results of information extraction are as follows: 6 tuple: [order, entity 1, entity 2, prefix, suffix, middle] = [1, improvement of theater facilities in various regions, box office, in recent years, significantly increased, led to].

After obtaining the tuple list, we can mine the list information. Firstly, due to the complexity of Chinese expression, in order to obtain the key morphemes, it is necessary to remove irrelevant parts, such as connection and pause from the text. The entity part of the meaning carrier is fixed and does not need to be processed. That is to say, this step mainly deals with prefix, suffix and middle. Secondly, after simplification, the samples are classified according to the order, and the samples of different orders are mined under the same set. Based on above samples, the following modes can be obtained:

Pattern 1: [because of, entity 1, correspondingly, entity 2, increased significantly];

Pattern 2: [in recent years, entity 1, lead to, entity 2, significantly increased];

Using pattern 1 and pattern 2 to extract the relationship of other texts, we can get other seed data satisfying the set relationship. And then put the new seed data into the seed set to get the pattern of the new seed, we can gradually expand the relationship data. When the relationship data, or other conditions meet the termination conditions, such as no data to be mined, the whole process can be terminated.

Through the extraction of the tuple relation in above two statements, the application of DIPRE algorithm is simply explained. As for the discourse relations discussed in this paper, the tuple information that needed to be extracted is not only limited to individual sentences; we also need to consider adjacent clauses. Therefore, the determination of tuple information is more complex. Generally speaking, to extract the relation between two adjacent clauses [sentence 1, sentence 2], the entity in the given seed should come from the two statements, and its tuple information should be a seven tuple: [order, prefix-e 1, entity 1, suffix-e 1, prefix-e 2, entity 2, suffix-e 2]. The order represents the order relationship between entity 1 and entity 2 in the statement. The entity 1 represents the entity from [sentence 1] in the relation seed; entity 2 represents the entity from [sentence 2] in the relation seed. Prefix-e1 represents the content immediately before entity 1; suffix-e 1 represents the content immediately after entity 1. Prefix-e2 represents the content immediately before entity 2; suffix-e 2 represents the content immediately before entity 2. The following sentences are taken as the example.

Sentence 1: he is ill, and he does not come to the class today.

Sentence 2: your body has not recovered yet, so you do not need to go to work.

The above two sentences are composed of two clauses, and their coherence is implicit causality. In order to extract the coherent relation, the relation seeds of two sentences are given respectively.

Sentence 1 seed: (sick, come to class);

Sentence 2 seed: (body, go to work).

According to the 7 tuple model proposed above, we can extract the information of the two statements respectively, and get:

Statement 1: 7 tuple: [order, prefix-e 1, entity 1, suffix-e 1, prefix-e 2, entity 2, suffix-e 2] = [1, he, sick, is, does not, come to class, -];

Statement 2: 7 tuple: [order, prefix-e 1, entity 1, suffix-e 1, prefix-e 2, entity 2, suffix-e 2] = [1, you, body, has not recovered, do not need to, come to work, have];

The model is as follows:

Pattern 1: [he, entity 1, is, no, entity 2, X];

Pattern 2: [you, entity 1, not recovered yet, no, entity 2, have].

It should be noted that in the information extraction result of statement 1, there is no suffix-e 2, resulting in an uncertain suffix "X" in the obtained pattern 1. In the practical application, "X" can be taken as any value, that is, when using pattern 1 in relation extraction of other texts, the content of suffix-e 2 in the statement can be ignored.

4. The Development of Dipre Algorithm

With the exponential growth of information on the Internet, the structure of web pages is becoming more and more diversified. There are limitations in the breadth and accuracy of data extracted by the DIPRE algorithm. Based on the concept of DIPRE, a "snowball" algorithm is developed to obtain structured data information from plain text documents with the minimum manpower.

The Snowball pattern develops key components of the DIPRE algorithm, and proposes a new technology to generate patterns and extract tuple from text documents. The algorithm also introduces a strategy to evaluate the quality of patterns and tuple information generated iteratively in each extraction, only tuple and patterns with high confidence coefficients are considered. This generation and filtering strategy significantly improves the quality of extracted relations.

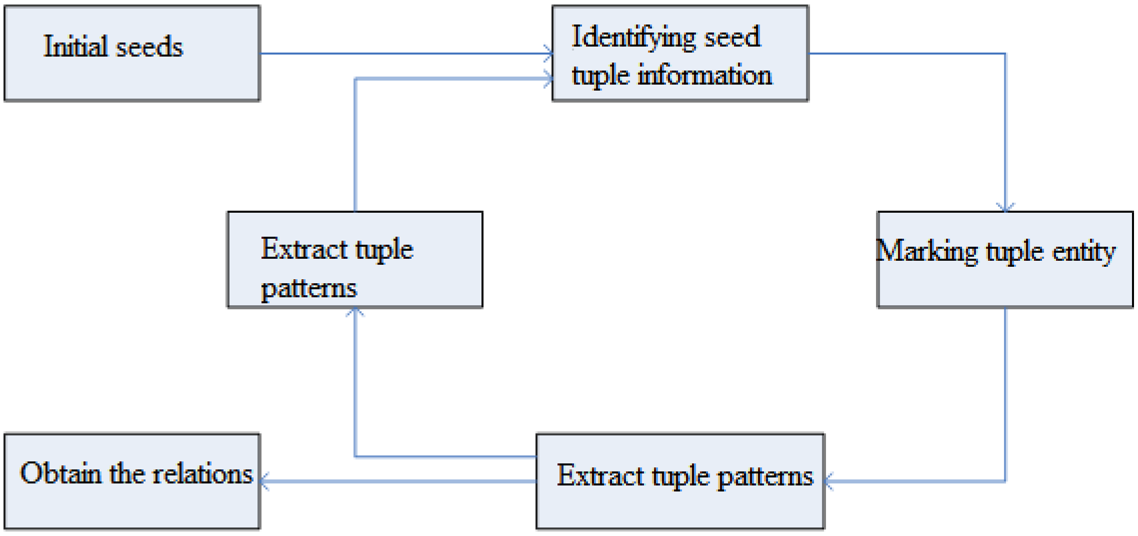

Fig.1 The Calculation Framework of Snowball.

A key step in the relation extraction process is, the generated pattern is used to find new tuple in the document. Ideally, we want patterns to be selective, that is, to be determined in advance, so that they don't generate bad tuple and have high coverage to recognize more new tuple. The Snowball first gives some sample tuple. Compared with DIPRE, the main improvement is to mark tuple entities. The key step in generating and subsequent matching patterns is to find where entities appear in the text. At the same time, it ignores irrelevant information. Some minor changes, such as extra commas or qualifiers, do not affect the extraction of relationships. After the pattern is generated, the Snowball finds new tuple through continuously gathering information. Its goal is to extract as much valid tuple as possible from the text set and merge them into a relational set.

5. Conclusion

Since the existing research results can not be directly applied to the automatic assessment of coherence relations in Chinese discourse, researches on the automatic assessment of discourse coherence based on the large-scale text corpus is an urgent issue. Starting from the DIPRE algorithm, this paper makes a preliminary exploration on this problem and proposes a method to extract discourse coherence relations, especially the implicit coherence relations. The research shows that the snowball algorithm developed by DIPRE and its application are of practical values for the automatic assessment of discourse coherence relations in large-scale text corpus.

References

[1] Agichtein, E. and Gravano, L. (2000). Snowball: Extracting Relations from Large Plain-Text Collections, Proc Acm DI.

[2] Feng, Z.W. (2008). Formal Model of Natural Language Processing, University of Science and Technology of China Press.

[3] Zhang, M.Y., Song, Y., Qin, B., et. al. (2013). Chinese Discourse Relation Recognition. Journal of Chinese Information Processing, no. 6, pp. 51-57.

[4] Zhang, M.Y., Qin, B. and Liu, T. (2014). Chinese Discourse Relation Semantic Taxonomy and Annotation. Journal of Chinese Information Processing, no. 2, pp. 28-35.

[5] Li, Z.W. and Liang, G.J. (2018). On the Computability of Discourse Coherence. Foreign Languages Research, no. 2, pp. 27-31.