2014

ISSN : 0974 - 7435 Volume 10 Issue 20

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(20), 2014 [12439-12444]

Mathematical model for Chinese sentence clustering based on dependency syntax

Xiaoqi Niu[1], Jing Xiong[2,*]

[1]School of Civil Engineering and Architecture, Anyang Normal University, Anyang 455000, (CHINA)

[2]School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, (CHINA)

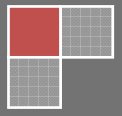
E-mail : xionghb125@sohu.com

ABSTRACT

This paper uses multi-feature fusion method to research and analysis the Chinese dependency syntax tree. A similarity computing model for dependency syntactic tree is proposed. Based on the dependency syntactic tree structure, the node words, parts of speech, and the dependencies between words are considered. The similarity calculation method between two dependency syntactic trees is proposed through comprehensive analysis of feature weights of dependency relation. The experimental results show that this method achieved a high accuracy rate.

KEYWORDS

Chinese information processing; Dependency syntax; Similarity calculation; Sentence clustering; Relative range.



INTRODUCTION

Sentence clustering is an important part in natural language processing. It provides an effective way for information extraction, subject classification and automatic abstract generation. Sentence clustering methods for Chinese information processing are the method based on feature words[1], the method based on iterative self organizing algorithm[2], the method based on multi feature[3], and so on.

During sentence clustering, the sentence similarity computation is an important basic. Currently, the research of sentence similarity computation can be roughly divided into some kinds as follows: words-based similarity computation[4, 5], similarity computation based on word semantic[6], similarity computation based on syntactic structure[7-9], similarity computation based on edit distance, and similarity computation based on dynamic programming[6]. Among these methods, the first three need sentence segmentation processing and the latter two do not. In recent years, the valence theory and dependency grammar syntactic structure analysis have become research hotspots. Syntactic structure similarity computing lays the foundation for the sentence similarity computation, so it important and necessary to study the syntactic structure similarity computation based on dependency relations.

This paper presents a mathematical model for calculating the dependency tree similarity, the model not only pay attention to the words themselves, but also study on the POS and dependency relations between words. By using the similarity calculation of two syntactic dependency trees we can implement sentence clustering.

OVERVIEW OF DEPENDENCY GRAMMAR

Dependency grammar is first put forward by the French linguist L. Tesniere in his book Elements of Structural Syntax, published posthumously in 1959. It will sentence analysis into a dependency tree, describe the dependencies between words. It is also pointed out that in the syntactic collocation relations between words, such collocation is associated with semantic. Dependency grammar has a far-reaching influence on the development of linguistics, especially in the area of computational linguistics respected. It uses the analysis of the relationship between syntactic dependency language units within the composition to reveal its syntactic structure. It advocates that center on verbs, verb dominating the other components, but they themselves are not dominated by any other component, all controlled by the components in some dependency from belongs to the dominant[7,10].

There are some common methods to parse dependency syntax, such as rule-based, statistics-based, combined method based on rules and statistics, tree library based[10]. For Chinese research the Treebank include U-Penn tree library built by University of Pennsylvania[11], Sinica Treebank of Central Research Institute of Taipei[12], Chinese syntax Treebank of Tsinghua University[13] and Harbin Institute of Technology dependency Treebank[14]. Dependency parsing expression form is divided into tree structure and relation of set form. An example of Chinese sentence tree structure is shown in Figure 1.

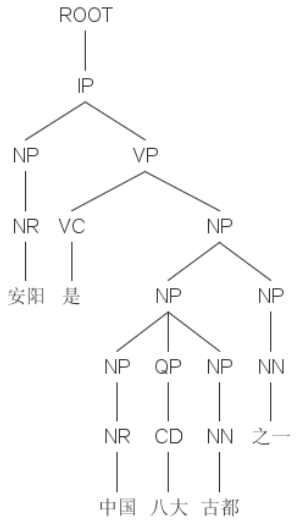


Figure 1 : Sample of Chinese dependency syntax tree

Dependency syntax is widely applied in the field of Natural Language Processing. Hu et al.[15] studied syntax path similarity calculation based on the dependency syntactic data, and realized the answer extraction of Chinese question answering system. Li et al.[16] exploited the dependency tree for case comparison to detect multiple constraint relations within a sentence, collecting signal words and relevant examples to construct a case-base of different kinds of constraint relations; defined a Partial Dependency Tree kernel for similarity measurement. Gu et al.[17] proposed an approach for extraction based

on semantic dependency in order to deal with Chinese ontology non-taxonomic relation extraction. They used semantic role labeling and dependency grammatical analysis to obtain semantic role and dependency relation of text sentence element. Verb frame with semantic dependency from text was extracted, and then discovered the non-taxonomic relation between concepts and name of the relation by computing semantic similarity. Most of the research and application based on dependency grammar are considered in advantages of concise form and easy to label and application[18].

SIMILARITY COMPUTATION MODEL OF DEPENDENCY SYNTACTIC TREE

Definition 1: To define a dependency relation R (C, A, D, CP, AP) as a five tuple, where C is the core word, A is the dependent word, D is the dependency type, CP is the POS of word C and AP is the POS of word A. The elements of R are characteristic parameters and each one of them has different feature weight.

According to the dependency theory, one node or word in dependency syntax tree can only has one dominating node or word but can have multiple dependent nodes or words. Thus in R tuple the C is more important than the A. In addition, a word may have different parts of speech, and each part of speech contains many words, so the importance of the words themselves is heavier than their POS. Finally, the dependency relation D between two certain words not only related to the words themselves but also related to their POS, so the importance of D is between the word itself and the part of speech. Therefore the feature weights order of the elements in dependence relation R is: C>A>D>CP>AP.

Under the assumption that there is a dependency relation pair $<R_1, R_2>$, to contrast the 5 elements in R_1 (C_1, A_1, D_1, CP_1, AP_1) and R_2 (C_2, A_2, D_2, CP_2, AP_2) respectively, if the two feature parameters are equal the value is 1 and the value is 0 when they are not equal. Then, according to the order from high to low weight arranged the five 0 or 1 values, get a binary number $(bbbbb)_2$. The range of the binary number is between 0 and 31, which corresponds to 0 means R_1 and R_2 completely unequal, and 31 corresponding to the situation that R_1 and R_2 are exactly the same. Based on the binary number, the similarity definition of R_1 and R_2 are as follows:

$$R_1 \mid R_2 = S(R_1, R_2) = \frac{(bbbbb)_2}{(11111)_2} \times 100\% \quad \textbf{(1)}$$

For example, there are two five tuple R_1 (C_1, A_1, D_1, CP_1, AP_1) and R_2 (C_2, A_2, D_2, CP_2, AP_2), assumption the compare results are $C_1=C_2$, $A_1\neq A_2$, $D_1=D_2$, $CP_1=CP_2$ and $AP_1\neq AP_2$, then according to the feature weight order that the binary number is (10110) 2. So the similarity between R_1 and R_2 is as follow:

$$R_1 \mid R_2 = S(R_1, R_2) = \frac{(10110)_2}{(11111)_2} \times 100\% = \frac{22}{31} \times 100\% = 70.97\% \quad \textbf{(2)}$$

Suppose there are two dependency relation pair sets $A=(a_1,a_2 \ldots a_n)$ and $B=(b_1,b_2 \ldots b_m)$, without loss of generality, let the element number of the set A is less than the element number of set B. That is, $n \leq m$.

In order to compute the similarity of set A and B, it needs to determine the corresponding dependency relationship belongs to set A and set B respectively. For each $a_i \in A, 1 \leq i \leq n$, there are several corresponding relations: $b_j \in B, 1 \leq j \leq m$. Suppose different a_i corresponding to different b_j, so the total number of corresponding relations about set A and set B is calculated as follow:

$$n \times (n-1) \times (n-2) \times \ldots \times (m-n) = \frac{n!}{(m-n)!} \quad \textbf{(3)}$$

Definition 2: if there is correspondence relation between set A and set B named Ω_k $(1 \leq k \leq \frac{n!}{(m-n)!})$, for a given a_i in Ω_k there is a b_j matching to a_i, denoted as $b_j = \Omega_k(a_i)$. The similarity of Ω_k is defined as:

$$Sim(\Omega_k) = \frac{\sum_{i=1}^{n} S(a_i, \Omega_k(a_i))}{m} \quad \textbf{(4)}$$

Definition 3: The similarity of the two dependency relation pair sets A and B is the maximum similarity of Ω_k, that is:

$$Sim(A,B)=Max\{Sim(\Omega_k)\}=\frac{\sum_{i=1}^{n}Max\{S(a_i,\Omega_k(a_i))\}}{m},\ n\le m, 1\le k\le\frac{n!}{(m-n)!} \quad (5)$$

Now take two Chinese sentences as analysis example. The sentences are parsed by LTP[19] (Language Technology Platform) which developed by Harbin Institute of Social Computing and Information Retrieval Research Center. The dependency relations of each sentence are as shown in Figure 2.

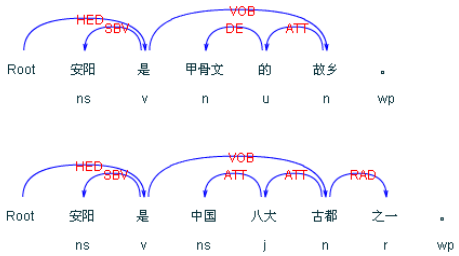


Figure 2 : Sample of dependency relations from two sentences

The corresponding dependency relation pairs illustrated in Figure 2 are shown in TABLE 1, which ignored HED relationship from Root. The $a_i|b_j$ (i and j can be equal) means the similarity between the two dependence relationship a_i and b_j, according to the formula (1) we can calculate the $a_i|b_j$ and choose the maximum one as the optimal value of $S(R_1,R_2)$.

TABLE 1 : Relation pairs of sentences

安阳是甲骨文的故乡(A)	安阳是中国八大古都之一(B)	$S(R_1,R_2)$ optimum value
a_1:SBV(是,安阳, SBV, v, ns) a_2:VOB(是,故乡, VOB, v, n) a_3:ATT(故乡,的, ATT, n, u) a_4:DE(的,甲骨文, DE, u, n)	b_1:SBV(是,安阳, SBV, v, ns) b_2:VOB(是,古都, VOB, v, n) b_3:ATT(古都,八大, ATT, n, j) b_4:ATT(八大,中国, ATT, j, ns) b_5:RAD(古都,之一, RAD, n, r)	$a_1\|b_1$= 100% $a_2\|b_2$= 74.19% $a_3\|b_3$= 19.35% $a_4\|b_2$= 3.23%

Thus, by using formula (5) the similarity of dependency syntactic tree can be calculated as follows:

$$Sim(A,B)=\frac{\sum_{i=1}^{n}Max\{S(a_i,\Omega_k(a_i))\}}{m}=\frac{1+0.7419+0.1935+0.0323}{5}=39.35\% \quad (6)$$

EXPERIMENT RESULT AND DISSCUSS

Based on the comparison of similarity of two sentences, we can complete the sentence clustering. In our experiments each sentence should be computed the similarities with other sentences, and take corresponding sentence which lead to the similarity maximum for the candidate sentence. After the completion of all sentences comparison each other, the clustering operation can be implemented. If there are two sentences mutually candidate sentences (described as 1:1 relation), they belong to the same class; if the sentences are described 1: n or n: 1 or m: n relation, then calculate the relative range value of their similarity to determine the result. Only when the relative range is less than a certain threshold d, the sentences can be classified as the same class. A sample of sentence clustering is shown in TABLE 2.

TABLE 2 : Sentence clustering sample

sentence number	experimental sentences	candidate sentences (sentence number)	similarity (%)
1	安阳是甲骨文的故乡	安阳是中国八大古都之一 (2)	39.35
2	安阳是中国八大古都之一	安阳是甲骨文的故乡 (1)	39.35
3	中国文字博物馆坐落在古都安阳	安阳是中国八大古都之一 (2)	20.97
4	洛阳有十三朝古都之称	南京素有六朝古都之称 (9)	44.09
5	周易的诞生地就是现在的古都安阳	甲骨文是全球最大的数据库软件公司(8)	56.45

6	周易分为易经和易传两个方面	周易的诞生地就是现在的古都安阳(5)	12.90
7	甲骨文是现存中国最古老的一种成熟文字	周易的诞生地就是现在的古都安阳(5)	44.84
8	甲骨文是全球最大的数据库软件公司	周易的诞生地就是现在的古都安阳(5)	56.45
9	南京素有六朝古都之称	洛阳有十三朝古都之称 (4)	44.09
10	甲骨文是王室用于占卜记事而在龟甲或兽骨上契刻的文字	甲骨文是现存中国最古老的一种成熟文字 (7)	43.32

From the experimental results shown in TABLE 2, sentence 4 and sentence 9 are just the one to one relation (denoted with $4 \leftrightarrow 9$), so they belong to the same class. And the other sentences should be judged. Now take sentence 1, 2 and 3 as an example to explain the judge process (the relation of the 3 sentence can be denoted as $3 \rightarrow 2 \leftrightarrow 1$).

Suppose that the similarity threshold d=15%, the average similarity of $3 \rightarrow 2 \leftrightarrow 1$ is $\overline{S} = 33.22\%$, and the range is $R = 18.38\%$, so the relative range is $RR = \frac{R}{\overline{S}} = \frac{18.38\%}{33.22\%} = 55.32\% > 15\%$ which beyond the threshold range. Therefore they could not be divided into one class. After analysis and eliminate sentence 3, sentence 1 and sentence 2 finally be classified as one class. The clustering processes of other sentences are similar, do not describe in detail.

In order to detect the effect of our clustering algorithm, we selected 200 sentences to experimentalize. The 200 sentence has been artificial selection and classification, involving 10 kinds of topics, including education, politics, sports, tourism, computer and other fields, and each kind of topics containing 18-22 sentences. The correctness of the experimental results only based on the similarity of clustering sentences, and do not consider the correctness of clustered categories. All segmented words of sentences were selected in the experiment, and the feature words were not screened. The experimental results are as shown in TABLE 3.

TABLE 3 : Sentence clustering contrast experiment

Method	Precision (%)	Recall (%)	F_1 (%)
literature 3	66.9	60.2	63.4
our method	72.2	55.4	62.7

From TABLE 3 it can be seen that our clustering algorithm increases slightly in accuracy compared to literature 3, but decreases in the recall rate. The main reasons are: in the accuracy, the dominate words, POS and dependency relation are all taken into account in our method. In addition, the accuracy rate is directly related to the results of dependency syntax analysis; in the recall, our method does not consider the semantic relationship between words, and there are only 0 and 1 cases about similar attributes of words, and thus leads to lower recall rate.

CONCLUSIONS

This paper presents a sentence clustering method based on the similarity of dependency syntax tree. The method is based on the dependency relation pairs, the dependency relation pairs set belong to different sentences will be matched and calculated the similarity. Selecting the correspondences which cause the sum of dependency relation pairs maximal as the computing fundament, and finding the average of the sum of dependency relation pairs similarity. It will be considered as the syntactic structure similarity of two sentences. The algorithm compares five features: lexical itself, part of speech and dependency relations, comprehensive measure of the similar relationship between the syntactic structures of sentences. But the present algorithm is lack of semantic analysis, and does not consider the influence of stop words when computing the similarity. In the next step, we will use the semantic advantages of ontology to perfect the algorithm. The influence of stop words and the weight factors of different dependency relation type and POS will also be considered.

ACKNOWLEDGEMENT

This research is supported by Development Projects of Henan Province Science and Technology (No. 132102210264), the Science and Technology Key Project of Henan Province Education Department (No. 14A520038) and Shandong Province Science and Technology Development Plan (No. 2013GGX10127).

REFERENCES

[1] Jian Xiao-Yan; Sentence Clustering of Farm Crops Based on Feature Words, Journal of Taiyuan Normal University (Natural Science), **1**, 77-79 **(2008)**.

[2] Fang Ying; Algorithm of the Sentence Clustering Based on Feature Selection, Modern Computer, **5**, 23-25 **(2007)**.

[3] Fang Ying, Yang Er-Hong; Research on Sentence Clustering Methods Based on Multi-feature, JSCL, 369-374 **(2005)**.
[4] Yang Si-Chun; An Improved Model for Sentence Similarity Computing. Journal of University of Electronic Science and Technology of China, **6**, 956-959 **(2006)**.
[5] Kahaerjiang Abiderexiti, Tuergen Yibulayin, Yao Tianfang, et al; An ImprovedMethod for Uyghur Sentence Similarity Computation. Journal of Chinese Information Processing, **4**, 50-53 **(2011)**.
[6] Liu Hong-zhe. Ontology Based Sentence Similarity Measurement. Computer Science, **1,** 251-256 **(2003)**.
[7] Li Bin, Liu Ting, Qin Bing, et al; Chinese Sentence Similarity Computing Based on Semantic Dependency Relationship Analysis. Application Research of Computers, **12**, 15-17 **(2003)**.
[8] Liu Bao-yan, Lin Hong-fei, Zhao Jing. Chinese Sentence Similarity Computing Based on Improved Edit-Distance and Dependency Grammar, Computer Applications and Software, **7**, 33-34 **(2008)**.
[9] Feng Kai, Wang Xiao-hua, Chen Zhi-qun; Chinese Sentence Similarity Algorithm Based on Dynamic Programming. Computer Engineering, **2**, 220-224 **(2013)**.
[10] Liu Hai-Tao; Dependency Grammar from Theory to Practice. Beijing, Science Press, **(2009)**.
[11] N.Xue, F.Xia, F.Chiou, et al; The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural language engineering, **11(2)**, 207-238 **(2005)**.
[12] K.Chen C.Luo, M.Chang, et al; Sinica treebank: Design criteria, representational issues and implementation, 231-248 **(2003)**.
[13] Zhou Qiang; Annotation Scheme for Chinese Treebank. Journal of Chinese Information Processing, **18(4)**, 1-8 **(2004)**.
[14] T.Liu, J.Ma, S.Li; Building a Dependency Treebank for Improving Chinese Parser.. Journal of Chinese Language and Computing, **16(4)**, 207-224 **(2006)**.
[15] Hu Bao-Shun, Wang Da-Ling, Yu Ge, et al; An Answer Extraction Algorithm Based on Syntax Structure Feature Parsing and Classif ication. Chinese Journal of Computers, **31(4)**, 662-676 **(2008)**.
[16] Li Huan, Liu Wen-Yin, Chen Xiao-Ping, et al; Exploiting Dependency Tree forMultiple ConstraintRelation Detection. Journal of Chinese Computer Systems, **31(6)**, 1112-1116 **(2010)**.
[17] Gu Ling-Lan, Sun Su-Yun; Approach to Chinese ontology non-taxonomic relation extraction. Computer engineering and design, **33(4)**, 1676-1681 **(2012)**.
[18] Ma Jin-Shan; Research on Chinese Dependency Parsing Based on Statistical Methods. Harbin: Harbin Institute of Technology, **(2007)**.
[19] W.Che, Z.Li, T.Liu; Ltp: A Chinese language technology platform. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, **(2010)**.