Research on the Removing Overlapping Ambiguity of Chinese Segmentation based on the Granular Computing model of Ontology

Fanjin Mai [a], Gen Zhang [b]

Guilin University of Technology, Guilin 541006,China

[a]gltide2010@126.com, [b]389978008@qq.com

Keywords: Concept, Ontology, Granular Computing, Difference of t- test, overlapping ambiguity.

Abstract. The word as the smallest language unit has its own independent concept. Human beings learn language through understanding semantics. This paper proposed a Granular Computing model based on Ontology, to restrict the concept of word and word, the use of Granular Computing similar ideas as well as human stratification and difference of t-test statistics method, the paper mainly studies the Chinese segmentation in overlapping ambiguity. The feasibility and effectiveness of the proposed model and the computational method are demonstrated by simulation experiments.

1. Introduction

Unlike the written language of the west, there are no boundaries between the words and the word of the Chinese language. Therefore, the problem of automatic segmentation of Chinese language becomes the primary work in computer processing Chinese [1]. There are two main problems in word segmentation: The first is the segmentation problem of ambiguity, the ambiguous segmentation is divided into overlapping ambiguity and combinational ambiguity; Second is the problem of unknown word [1].The segmentation of ambiguous is the important factor to affect accuracy and speed of segmentation system. So it is very important to propose or improve the method of eliminating the ambiguity.

This paper attempts to use a hierarchical thinking pattern similar to humanity, and the granular computing model based on Ontology is introduced. Focus on the surge of Chinese information in the environment of today's large data. Using the massive corpus training and the difference of t- test statistical method, the overlapping ambiguity is distinguished. And then, I want to develop a more similar Chinese word segmentation method.

2. Disambiguation of Chinese Word Segmentation

2.1 Generation and Recognition of Ambiguity.

The reason for the ambiguity is that there is no boundary mark between the words and the words in the Chinese text, so it looks just a sentence for string. In addition, the ability of the Chinese morpheme to construct vocabulary and the multi-function of the category of the Chinese word and there are a large number of place names are the reasons for the emergence and increasing ambiguities [2].

The main forms of the ambiguous segmentation are the overlapping ambiguity and the combinational ambiguity. A bidirectional maximum matching is generally adopted for identifying the overlapping ambiguity. The main words of the combination ambiguity are based on the dictionary of the word segmentation or the combination ambiguity is identified by the way of establishing the combination ambiguous database.

According to the data statistics, the overlapping ambiguity in all the ambiguities accounted for about 90%. So the research is also focused on the overlapping ambiguity processing. In this paper, the focus of dealing with the overlapping ambiguity and the text uses the bidirectional maximum matching method for the initial segmentation and identifies the overlapping ambiguity field.

2.2 Disambiguation Method.

These algorithms can be classified into the following categories: the disambiguation method based on rule, the disambiguation method based on probability statistics and the disambiguation method based on the knowledge base and so on. In recent years, the method of probability statistics is relatively popular and mutual information (MI) and Hidden Markov model (HMM) are often used.

The definition of MI algorithm: For the two character A and B, the MI value M (A, B) is calculated by MI formula, thus the correlation degree between B and A is judged. The small value of MI is eliminated as ambiguity. The method of MI is simple and effective, but has nothing to do with context and the accuracy rate of disambiguation is not high

In a language model, the probability of a word depends on the word in front of it, the probability that the N word occurs only with it before the N-1 words. This is a language of the N-1 MM. When the MM of the observed sequence is known and the actual state is unknown, it is called the HMM [8]. The use of HMM for disambiguation needs for large-scale corpus training. Although the disambiguation accuracy is relatively high, but does not apply to some of the requirements of the high speed environment segmentation

In this paper, the model of granular computing based on Ontology has good efficiency in large-scale data processing and it is similar to human's thinking of the hierarchical model. T-test is a kind of probability statistical method of combining context. The combination of the two methods is a new attempt, in the elimination of ambiguity in both the accuracy and speed.

3. Granular Computing Model Based on Ontology

Using granular computing to model the important is how to construct the granule and the appropriate discourse domain [4]. Ontology describes the concept of semantics through the concept of relationships [3]. So we can use the concept model of Ontology to express different granularity.

According to definitions 3.2 from the literature [5], a complete semi ordered lattice is formed in the "≤" relation. Thus obtains: $R_n \leq R_{n-1} \leq \ldots \leq R_1 \leq R_0$ this sequence, and this sequence is relative to a N tree. Therefore, it can be inferred that the granularity can be represented by the concept hierarchy tree. Using the concept hierarchy of Ontology classification system is applied to the representation of granularity. *Guarino* et al. Proposed the Ontology driven modeling method in the literature [6].

Definition 1: Suppose there is a semantic, all words belonging to the semantic recorded as meaning of aggregate $s(\varphi)$ and expressed as $s(\varphi)=\{x \in U，x \mid\approx \varphi\}$. Among them U representation discourse domain. $\mid\approx$ is a formula that can satisfy the sign. $s(\varphi)$ is called a granule, and the semantic φ is the source of the granule.

Definition 2: Set $L(s(\varphi))$ as the size of the granule. $L\big(s(\varphi)\big) = {card\big(s(\varphi)\big)}\big/_{card(U)}$.$\mathrm{card}\big(s(\varphi)\big)$ indicates the number of elements contained in the $s(\varphi)$;$\mathrm{card}(U)$ expressed as the total number of discourse domain elements.

According to definitions 1 and 2, the semantic dictionary of the concept hierarchy based on Ontology can be established by statistical method. The concept of the structure should follow the relationship between concepts of Ontology. Discourse domain is the total vocabulary of the semantic dictionary. Granule is a collection of words contained in the semantics. The hierarchical relationship between class and class defined in the Ontology need to be carried out after the establishment of the ontology.

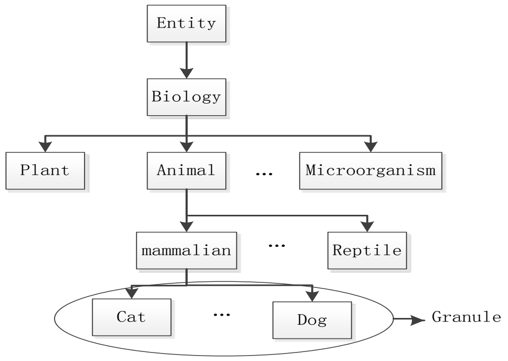

Fig.1 Concept hierarchy of biological domains

There are two methods to establish the hierarchy of classes in Ontology:

Top-down approach: Starting with the largest concept in a field, and then gradually adding a sub category to refine these concepts.

Bottom-up approach: Begin with the definition of the lowest and the smallest class, and then organize the detailed classes to be more comprehensive concept.

In this paper, the relationship between classes is established by using the top-down method. Concept hierarchy of biological domains, as shown in figure 1.

4. Eliminate Ambiguity

4.1 Disambiguation Model.

Experiments show that the accuracy of the word segmentation method based on statistics is higher than that based on rule, but its speed is slow [9]. In order to balance the accuracy and speed of word segmentation, we can use the method based on rule and statistics. Firstly, we use the bidirectional maximum matching method to segment the input text. And then compare the results of the two word segmentation, and found that the part of the overlapping ambiguity.According to the granular computing model based on ontology, the part of the ambiguity is marked by the source of the granule.Extracting ambiguity and the front and back granule of the ambiguous field.Finally, the difference of t- test is used to examine the ambiguous granule and the front and back granule in the field of ambiguity.Measure the degree of combination between granules, and then the ambiguity of the segmentation.The discovery in the process of unknown words in the dictionary.Gives the final result of word segmentation, and updates the results into the corpus for future statistical use.The whole process of disambiguation can be divided into 3 steps, namely, the discovery of ambiguity, the elimination of ambiguity and extraction of ambiguity. Schematic diagram of disambiguation model, as shown in figure 2.

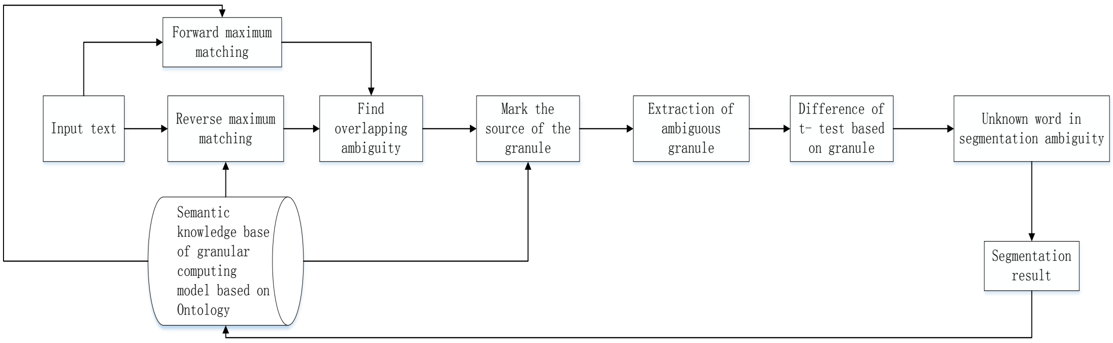
Forward maximum matching
Input text
Reverse maximum matching
Find overlapping ambiguity
Mark the source of the granule
Extraction of ambiguous granule
Difference of t- test based on granule
Unknown word in segmentation ambiguity
Semantic knowledge base of granular computing model based on Ontology
Segmentation result

Fig.2 Schematic diagram of disambiguation model

4.2 Calculation Method Based on Granule.

Church draws on the idea of T-test and T formula for calculating the degree of close between words is proposed. It is used to calculate a word in English with two other words, which is more closely related to one of them. However, the T-test is used to measure the associative degree of the comparative value, and cannot be compared to the context. Combined with t-test algorithm proposed by literature [7] and concept of granule, used to calculate the degree of association between two granules.

Definition 4: for granule x, y, z, the t-formula of y about x and z is defined as follows.

$$t_{x,z}(y) = {p(z \mid \mathrm{y}) - p(y \mid \mathrm{x})} \Big/ \sqrt{\sigma^2(p(z \mid \mathrm{y})) + \sigma^2(p(y \mid \mathrm{x}))}$$

The $p(z \mid \mathrm{y})$, $p(y \mid \mathrm{x})$ are z about y, y about x conditional probability, $\sigma^2\big(p(z \mid \mathrm{y})\big)$, $\sigma^2(p(y \mid \mathrm{x}))$ is the variance of their own.

From the definition of T–Test:(1) If $\mathrm{t}_{\mathrm{x,z}}(\mathrm{y})>0$, then the y tends to be connected with its posteriori z and break with x. The greater the value, the stronger the tendency. (2) If $\mathrm{t}_{\mathrm{x,z}}(\mathrm{y})<0$, then the y tends to be connected with x and break with z. Absolute value is more, the tendency is stronger. (3) If $t_{x,z}(y)$=0, there is no tendency.

Definition 5: the string of granules uxyw, the difference of t-test between X and Y is defined as follow.

$$\Delta t(x,y) = t_{u,y}(x) - t_{x,w}(y)$$

According to the definition 4, $\Delta t(x,y)$ is t-formula of x about u and y minus t-formula of y about x and w. $\Delta t$is a relative measure to the associative degree of x and y. But also relating to context (involve 4 granules).

5. Simulation Experiments and Results

5.1 Corpus training.

This paper contains two parts of the training corpus. The first part choose primary and junior high school Chinese teaching materials by the *People's Education Press* published from 2000 to now. Extract 674 texts from a total of 342573 words. The materials contain punctuation marks, which are not included in the ancient poetry, modern poetry, drama, classical Chinese and early vernacular. The second part choose *People's Daily* part of the text in July 2012. A total of 804722 words and this materials also contain punctuation marks. The total number of words is 1147295 from two parts.

5.2 Experimental results.

Experimental comparison of MI, HMM and the method is proposed in this paper. The identification of ambiguity is used in the bidirectional maximum matching method. From two aspects of accuracy rate and segmentation speed, experiments using the same training corpus. The effect of the model was evaluated by the accuracy of the word segmentation. The accuracy rate is calculated by the following formula:

$$\text{Accuracy Rate} = \frac{\text{The correct segmentation of words}}{\text{All the words segmentation}} \times 100\%$$

Experiment one and experiment two select the text of the two part of the *People's Daily* in November, 2012. The number of words were 12501 words and 9743 words. Experiment three choose not included junior middle school Chinese text *Emperor's New Clothe*s which a total of 2965 words. Test three kinds of disambiguation method.

Test results for the accuracy of the three methods, as shown in table 1.

Table 1 Test results

disambiguation method	Experiment 1	Experiment 2	Experiment 3
MI	91.56%	90.77%	89.86%
HMM	99.12%	98.96%	98.55%
My method	98.03%	97.85%	97.13%

Comparison of the segmentation speed of the experimental results with the three methods use bidirectional maximum matching method, as shown in figure 3

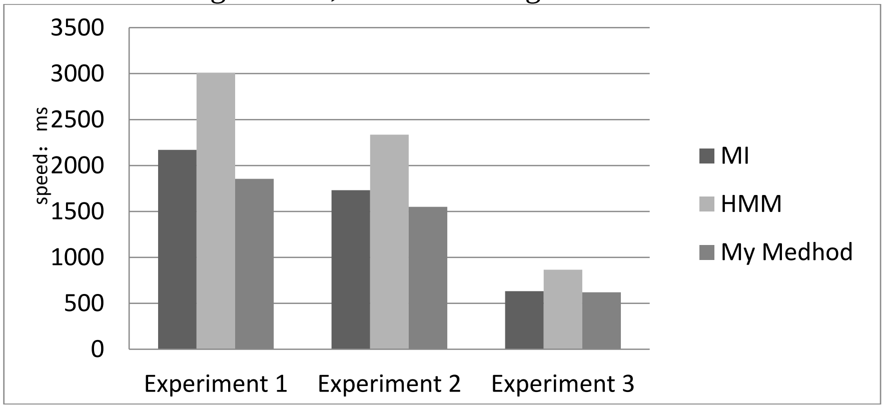

Fig.3 Comparison of word segmentation speed

According to the experimental results of table 1 and figure 3, it is not hard to see that the granular computing model based on Ontology has a better performance in the accuracy and speed of segmentation. However, the accuracy compared with HMM is slightly less, but there are obvious advantages in the segmentation speed. The comparison of MI is a better performance in all aspects. Due to the limited experimental conditions, we cannot carry out experiments on a larger corpus, but the author believes that this method will be more outstanding.

6. Conclusion

With the rapid development of the network, the amount of information is increasing in geometric multiples. Chinese word segmentation for large data volume of the text of the segmentation efficiency has higher requirements. The combination of Ontology method and granular computing. A model for the use of human's hierarchy and use the statistical method of difference of t- test to calculate the degree of the granule. Finally, the overlapping ambiguity field is segmented. Although there is no more large-scale corpus training, it has been proved by experiments that the method has good performance in the accuracy and speed of segmentation, and has high practical value.

Reference

[1] Chengqing Zong. Statistical natural language processing [M].Tsinghua University press.2013

[2] Yuzi Liu. Research on the algorithm of eliminating ambiguities in Chinese word segmentation [D]. Chongqing: Chongqing University.2005

[3] Zhihong Deng, Shi Tang, Zhangming Wei. Ontology Research Summary.Journal of Peking University. Vol. 38 (2002) No. 5, p.731.

[4] Ling Zhang, Ba Zhang. Problem solving theory and its application-The theory and application of quotient space granular computing [M]. The 2 edition. Tsinghua University press.2007

[5] Ling Zhang, Ba Zhang. Fuzzy quotient space theory [J]. Software Journal. Vol. 14 (2003) No. 4, p.770-776.

[6] Guarino N.Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. Inazienza MT, eds. Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Springer Verlag,1997 ,139～170.

[7] Maosong Sun, Changning Huang, Jiayan Zou, et al. Using the Chinese character two yuan grammar to solve the ambiguity of the overlapping type of Chinese automatic word segmentation [J]. Computer research and development. Vol. 34 (1997) No. 5, p.332-339.

[8] Fanjin Mai, Ting Wang. Word disambiguation model based on bidirectional maximum matching method and HMM [J]. Modern library and information technology. Vol. 2008 No. 8, p.37-41.

[9] Changning Huang, Hai Zhao. Chinese word segmentation ten years review [J].Chinese Journal of information. Vol. 21 (2007) No. 3, p.8-19.