The Research on Chinese Word Segmentation System with Semantic Annotations Information

Xian-Yi Cheng, Wei Kang, Jie Zhang and Quan Shi
Dept. of Computer Science and technology
Nantong University
Jiang Su, Nantong, China
e-mail: xycheng@ntu.edu.cn

***Abstract*—Chinese word segmentation technology is a very important basic work in the field of Chinese information processing. Aim to the existing Chinese word segmentation system only focus on grammar annotation, this paper will learning the rules of semantic annotation based on ID3 and feature selection based on CHI , the preliminary results of Chinese word segmentation for secondary annotations, the experiments show that enrich semantic information and advantageous to the named entity extraction by secondary annotations**

Keywords-Chinese; word segmentation; semantic annotations

I. INTRODUCTION

With the growing popularity of digital storage technology, the Chinese information resources is expansion at the speed of fast, it is hot and difficulty how to improve the efficiency of organization and management of Chinese information, and how to comprehensive, accurate and fast retrieval to the relevant information according to the needs of users in the field of Chinese information processing.

Because of the complexity of the Chinese language (large character set, continuous string), most of processing methods of western language cannot be transplanted into Chinese information processing directly. This is because, first of all, to the Chinese word segmentation, although has been a rapid growth of Chinese word segmentation technology, the Chinese word segmentation system there are still many problems to be solved[1].

For search engines, the most important thing is not to find any results, but arrange the most relevant results at the top. From a technical point of view, different segmentation algorithms and thesaurus will affect returned result in a retrieval.

Search engine is only an application of the Chinese word segmentation. Other applications include machine translation, speech synthesis, automatic classification, automatic summarization, automatic proofreading, relationship extraction so on, they all face the problems of quality of word segmentation in semantic level .To better serve more products the technology of the Chinese word segmentation has a long way to go[2].

In the existing Chinese word segmentation system, ICTCLAS, it is developed by China academy of sciences, regardless in the accuracy or speed have obvious advantages, but show some shortcomings of semantic depth in specific application. In this paper, the relationship extraction as the research object, to learn the rule of semantics by ID3 and select characteristics by CHI for the ICTCLAS segmentation result, and application learning rules of semantic annotation to ICTCLAS segmentation system again, experiments show that such secondary annotations, enriched the meaning of the word and is advantageous to information extraction, such as named entity and relationships.

II. THE RESEARCH STATUS

A. Dictionary based on mechanical word segment method

Dictionary based on mechanical word segment method, that is scan strings. If you find the same substrings and word strings, then matching. Such segmentations usually add some heuristic rules, such as “forward/ reverse maximum matching”, “long term priority” and other strategies. The advantages are fast algorithm, all O(n) time complexity, simple and its effect is ok. There are also disadvantages, dealing with ambiguity and unknown words is not good. Most of the existing word dictionary is a mechanical dictionary. Its internal entries are mechanical arranged according to the certain order. Dictionary is not reflecting the relationship between words. Dictionary function is simply to provide segmentation matching object. In promoting semantic analysis today, this dictionary is clearly not in line with our needs[3].

B. Chinese word segmentation based on statistical method

Chinese word segmentation based on statistical method, letting probability theory as the theoretical basis. The emergency of the combination string of Chinese characters in the context abstract into random process. In the specific implementation process, statistical based word segmentation method often using mutual information, N statistical model and t test theory principles to carry out the process of Chinese word segmentation[4]. Calculating the probability of the word appears through a variety of parameters in the process of word segmentation, the maximum probability as

© 2014. The authors - Published by Atlantis Press

the final result. However, statistical based segmentation method requires an existing training set or corpus preprocessing, investing a lot of human labor and therefore its sheer complexity and scale, the large amounts of data probability calculation in the process of word segmentation result high spatial and temporal complexity.

C. *Chinese word segmentation method based on knowledge*

Knowledge based segmentation method is an intelligent process, including: expert system method, neural network method and ontology based method.

Currently mature Chinese word segmentation system are considered to combine several different algorithms or use a variety of algorithms to deal with the problems in order to achieve better segmentation results[5].

III. SEGMENTATION SECONDARY TAGGING ALGORITHM

A. *The problems of ICTCLAS*

ACE relationship tagging corpus for example,fig.1 shows the segmentation primary result of ICTCLAS.

Fig.1. Segmentation Primary Result of ICTCLAS

Seen from figure 1, the red underlined is segmentation ambiguity. For example, the correct segmentation of “有/vyou 关于/p 益文/nr2” is “有关/p 于益文/nr”, the correct segmentation of “文化局/n 长龙/n 映台/nr2” is “文化局长/Title 龙映台/nr”, title is the title label, ambiguity often come from names and places. Also commonly used phrases can’t be separated, but preliminary segmentation can. Such as “特/ng 色/ng”, “刺/v 激/v”, “声/n 音/n 引/v 领/v” and so on. If considered from the semantic level, some part of speech should be amended semantic annotation, such as “明天/t 2/m 号/q 晚上/t”. “明天 2 号晚上/Timex2” want to be marked as timestamp, “大安/n 森林/n 公园/n” want to be marked facilities “大安森林公园/fac”. Of course, these problems can’t be completed in the segmentation system based on grammar. In the case of semantic-based segmentation system can’t meet the demands, the segmentation post-processing is necessary because the semantics are dependent on the specific applications.

B. *Semantic rules learning*

Semantic rules learning uses Weka ID3 algorithm[18], the process is as follows:

1) Constituted by the signature file. Generating the original characteristics for the different named entities(names, places, timestamp, organization, facilities, national, coins, etc). Three parts consist the feature: assuming the part of speech of current word is pos(0), before the word is pos(-1), after the word is pos(1), pos(i) constituted by the annotation norms of ICTCLAS. Getting 3*75=225 original features.

2) Characteristic selection based on CHI: Each named entity separate as a class and each pos(-1)pos(0)pos(1) as a document. As a result, named entity annotation problem become a document classification problem. The goal of feature selection is reducing the dimension, improving the performance of named entity annotation. So class, document, word are three elements of feature selection that must be considered. The influential way of feature selection is CHI statistical method , calculated as follows:

$$\chi^2(t_k,c_i)=\frac{N(P(t_k,c_i)P(\bar{t}_k,\bar{c}_i)-P(\bar{t}_k,c_i)P(t_k,\bar{c}_i))^2}{P(t_k)P(c_i)P(\bar{t}_k)P(\bar{c}_i)} \quad (1)$$

Where $p(t_k,c_i)$ represents document frequency belongs to class c_i which containing entry t_k. $p(\bar{t}_k,\bar{c}_i)$ represents document frequency not belongs to class c_i which containing entry t_k. $p(\bar{t}_k,c_i)$ represents document frequency belongs to class c_i which not containing entry t_k. $p(t_k,\bar{c}_i)$ represents document frequency not belongs to class c_i which not

containing entry t_k. N represents the total number of the document corpus.

Equation (3) only calculate a value of CHI corresponding to a particular class. It is the integration of the value of all classes of CHI corresponding to the value of CHI of all text data. There are two comprehensive strategies, one is a weighted average, the second is the maximum value.

CHI method has the following disadvantages: ①can't distinguish the degree of importance between positive and negative related characteristics. In CHI calculation formula, p(tk,ci) and p(tk,ci) are positive related degree, another two are negative. ② since CHI method is based on the assumption of distribution, if the distribution assumption between feature words and text category is broken, CHI method would be more inclined to choose low frequency characteristic word.

So, this paper introduce the class frequency, concentration and dispersion to modify CHI model. Class word frequency is defined as the frequency that a certain feature items appear in all texts in certain class. If the frequency is greater, indicating the ability of showing such texts is more stronger. CTFik shows the size of class word frequency:

$$CTF_{ik}=\sum_{j=1}^{|c_k|} tf_{ij} / TF_i \qquad (2)$$

Where tfij represents the frequency that the i-th feature item appears in the j-th document. TFi is the general features of the i-th feature word. |ck| is the number of documents of the k-th class. Introducing class word frequency can effectively exclude those unreliable low frequency feature items. Concerning about the relationship between class and word.

Concentration is defined as the reciprocal of the number of class that a certain feature item appears in all documents. The more the numbers of category of feature items, the worse the ability of representing. So we use the reciprocal of the number of categories to indicate concentration. The concentration is bigger, the stronger the ability of showing, and vice versa. bi represents the size.

$$b_i=Class_i/|C| \qquad (3)$$

Classi represents the numbers of category that the i-th feature item exist in all training set. |C| represents total numbers of category. Introducing concentration in order to reduce the ability of feature item that appears in all categories.

Dispersion is defined as the ratio between the numbers of document which appear in a certain class and total numbers of document of this class. The size of dispersion represent the size of document coverage that certain feature items appear in certain class. The greater dispersion shows the more the numbers of document of this feature item covered in this class. So the ability of showing the class is stronger, and vice versa.

$$c_{ik}=feature_{ik}/N \qquad (4)$$

Feature ik is the number of document that the i-th feature item appears in the k-th class. N is the training document review. Introducing dispersion in order to exclude feature items appear frequently only in the individual documents.

Obtaining feature selection cube calculation model according to the formula (2),(3),(4).

$$Cube_{ijk}=(tf_{ij}*CTF_{ik}*c_{ik})/b_i \qquad (5)$$

Depending on the feature selection method of class word frequency, excluding the feature items that there is small coverage in class but bigger in other classes through dispersion. Using concentration to exclude the feature items evenly distributed in all kinds of classes. Using class word frequency to exclude unreliable low frequency feature items.

According to the formula(5), the feature dimension can be reduced between 10 to 20.

generating semantic rules

Making the decision trees which is obtained by WekaID3 algorithm after the feature file reduced dimension as a candidate rules. Increasing necessary prior knowledge on this basis, such as "说、讲" and so on, there are two or three words in front, while first word is surnames, combining words as names. If the length of label "NR" is 2 and following by the word, merging words as three word names, etc. After obtained the semantic rules and prior knowledge, we use GATE Jape plug to analysis system, the segmentation result is shown in fig 2.

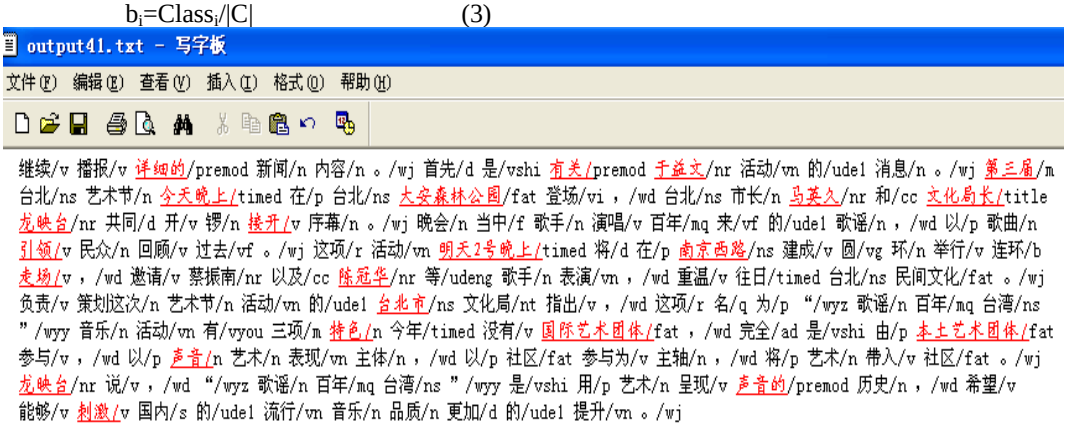

Fig. 2. Segmentation Result

Seen from fig.3, segmentation post-processing result can basically meet the demand of named entity annotation application. Containing rich semantic information and laying the foundation for follow deeper semantic annotation.

IV. EXPERIMENTAL ANALYSIS

A. Experimental data

1)January 1998 "People's Daily" annotated corpus

2)ACE2005 Chinese relations annotated corpus. The corpus have 578 articles, 458 articles are training corpus, 120 articles are test corpus Table 1 list the statistics of named entity in ACE2005 corpus[6].

TABLE1. THE NUMBEROF ALL ENTITY CATEGORIES

NE type	NE count Within the scope of relations	NE count
PER(Person Name)	3415	6050
ORG(Organization)	2147	3672
LOC(Location)	803	1016
FAC(facilities)	898	1024
GEP(Geography)	2449	3062
WEA(weapons)	213	234
VEH(vehicle)	324	365

The first step is preliminary segmentation, POS tagging and then generating "named entity-feature" matrix according to the feature selection algorithm given by section 3. On this basis, using ID3 algorithm in Weka to learn named entity semantic annotation rules, finally generate evaluation for new annotation result.

B. Experimental and analysis

Experiment 1: Analysis: the low precision and recall rate of ICTCLAS for "names(PER)" because evaluation criteria for PER is different from ACE and ICTCLAS, such as "singer, himself, owner, everybody, mayor and so on" defined as PER in ACE, but there is no such requirement in ICTCLAS. This article is based on ACE training corpus standard. Therefore, the accuracy rate PER marked can meet the basic requirement. Low recall because of the suffix title, professor, director and so on. False consciousness caused by insufficient training corpus.

TABLE II. COMPARING ICTCLAS SEGMENTATION WITH POST-PROCESSING

NE	Accuracy		recall rate	
	ICTCLAS	Our	ICTCLAS	Our
PER	38.42	87.31	20.58	85.78
ORG	82.17	90.44	85.28	88.12
LOC	92.53	93.74	95.26	90.55

Both methods of LOC are the same, the reason is ICTCLAS and ACE supports LOC label, which is also close to the standard.

Because ICTCLAS doesn't support the four standards (FAC、GPE、WEA、VEH), so the two methods are not comparable.

Description: in this experiment, low performance of ICTCLAS does not represent the low performance in other application. Because the purpose of this experiment is to

Extract relationships, in this application background the performance of word segmentation based on ACE might be better than ICTCLAS.

experiment 2:Literature[3] uses cascading condition random field model to identify named entity. The method uses underlying model to identify simple named entity, choosing several best results to deliver to senior condition random filed model. Table 3 gives the comparison experiment of the method of literature[3] .

TABLE 3. THE COMPARISON EXPERIMENT OF THE METHOD OF LITERATURE[3] WITH PROPOSED METHOD.

NE	Accuracy		recall rate	
	literature[3]	Our	literature[3]	Our
PER	85.23	87.31	88.31	85.78
ORG	79.62	90.44	80.48	88.12
LOC	80.91	93.74	81.43	90.55

he overall performance is better than the literature[3] because the text only mark named entities within the scope of the relationships not the entire document.

experiment 3:The experiment tests different corpus using text method(table 4).

TABLE 4. PERFORMANCE COMPARISON OF DIFFERENT CORPUS

NE	Accuracy		recall rate	
	People's Daily	ACE2005	People's Daily	ACE2005
PER	87.94	95.31	89.02	95.78
ORG	89.62	90.73	83.88	91..42
LOC	91.61	92.51	90.43	92.75

Named recognition of this paper has obvious advantages, because the ambiguity of names is much bigger than others, another reason is that the definition of names of two corpus is different. The method of this paper is designed for ACE. Although there is not comparable, but illustrating that the method is also applicable to other corpus.

V. CONCLUSION

Chinese as one of the most frequently used language in the world has its own characteristics. Chinese word segmentation technology is the basic work of Chinese information intelligent processing. An efficient and accurate Chinese word segmentation system can bring great convenience to other work of Chinese information processing. Semantic rules learning is presented in this paper based on the ICTCLAS segmentation. Enriching the semantic information of word segmentation, making for named entity extraction and relation extraction application.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of China National(No: 61340037,61202006).

REFERENCES

[1] Huang cn, Zhao h. Chinese Word Ten Years[J]. Journal of Chinese Information,2007,21(3):8-19

[2] Wang s, Cao cg, Pei yj. A Chinese Lexical Semantic Similarity Calculation Method Based on Collocation[J]. Journal of Chinese Information, 2013,27(1):7-15.

[3] Tan wx. Identification Study on Named Entities and Basic Noun Phrase[D]. Suzhou, Suzhou University Master's Thesis. 2010

[4] Zhao y. The Study on Level Chinese Segmentation Method Based on the Text Classification[D]. Guilin, Guangxi University Master's Thesis,2012.

[5] Qin w, Yuan cf. Chinese Unknown Word Recognition Based on Decision Trees[J]. Journal of Chinese Information, 2004, 18(1):14-19.

[6] http://www.freeloongson.com/.