A Chinese Word Segmentation System Based on Cascade Model

Zhang Jianfeng
School of Computer and Information Technology, Shanxi University
Zhangjianfeng83@163.com

Zheng Jiaheng
School of Computer and Information Technology, Shanxi University
jhzheng@sxu.edu.cn

Zhang Hu
School of Computer and Information Technology, Shanxi University
zhanghu@sxu.edu.cn

Tan Hongye
School of Computer and Information Technology, Shanxi University
Hytan_2006@126.com

Abstract

This paper introduces the system of Word Segmentation and analyzes its evaluation results in the Fourth SIGHAN Bakeoff[1]. A novel method has been used in the system, which main idea is: firstly, the main problems of WS have been classified, and then a cascaded model has been used to gradually optimize the system. The core of this WS system is the segmentation of ambiguous words and the internal information extraction of unknown words. The experiments show that the performance is satisfying, with the RIV-measure 96.8% in NCC open test in the SIGHAN bakeoff 2007.

1 Introduction

Chinese Word Segmentation is a fundamental task for some Chinese NLP tasks, such as ma chine translation, speech recognition and information retrieval etc. However, the current performance of WS is not satisfying. In WS the disambiguation processing and unknown words recognition are the two difficult problems. So, we aim at the solution of the both problem in our WS system. We participated the SIGHAN bakeoff 2007 evaluation, and a cascade model has been used in the process of word segmentation. In the WS system, the core modules are the segmentation of ambiguous words and the extraction of internal information of unknown words.

2 System Description Introduction

Figure1 shows the workflow of our WS system. The system is made up of the following modules: small sentences segmentation, disambiguation, and unknown words recognition.

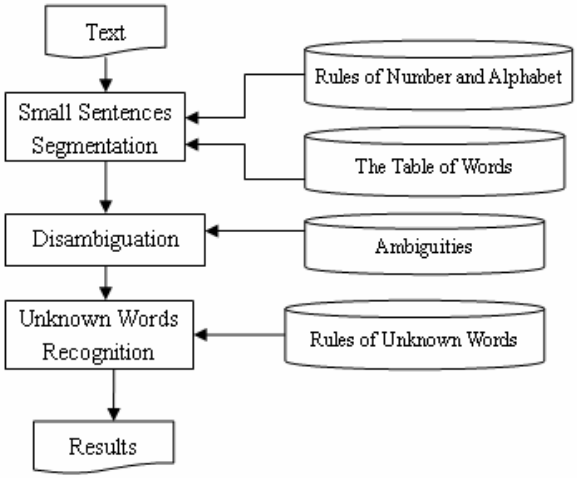

Figure 1 Segmentation System based on cascade model

[1] This research was partially supported by the National Natural Science Foundation of China No. 60473139, the National Natural Science Foundation of China No. 60775041and the Natural Science Foundation of Shanxi Province No. 20051034.

In fact, ambiguities may appear between two lexical words or between a lexical word and an unknown word. Based on the observation, the disambiguation processing is prior to the unknown word recognition in the system.

2.1 Disambiguation

We classify the ambiguities into two categories: one is true-ambiguity, the other is pseudo-ambiguity. And the process of the true-ambiguity is the disambiguation emphasis.

According to the NCC training corpus, we build pseudo-ambiguity database. For pseudo-ambiguity, disambiguation can be realized through matching against the database.

We get all ambiguities from the training corpus. Pseudo-ambiguity can be solved by finding the database of pseudo-ambiguities which built on the base of the analyses of NCC corpus.

For true-ambiguities, we also build a database and use a statistical model to disambiguate.

Based on the examination of the database of ambiguities, we find that true-ambiguities appear in the two cases: （1）both frequencies of two segmentations of ambiguities are low, or the gap between the two frequencies is too large; (2) both frequencies of two segmentations of ambiguities are high. For the former case, the segmentation form corresponding to the lower frequency is saved in the database. For the latter case, both segmentations and their context are saved in the database. And the system will choose the appropriate segmentation according to the statistic model.

The statistic model can be represented as the following formulas:

$$y = \arg\max_{y} p(y \mid x)$$

$$p(y \mid x) = \sum_{i=0}^{3} bi \sum_{j} f_{ij}(x, y)\, p(x, y)$$

$$p(x, y) = \frac{freq(x, y)}{\sum_{x \in X, y \in Y} freq(x, y)}$$

Among the formulas, x is the context, and y is the segmentation form, fi(x,y) is the feature functions, p(x,y) is the empirical probability, and bi is the impact factor of the feature function, whose value is determined according to the Tongyici Cilin[2]. Here, (x,y) can considers not only the neighboring words but also the semantic information of the neighboring words.

The impact factor bi is defined as follows:

Let $p \in pre(S), t \in next(S)$, so

$$bi = \begin{cases} 2, & p \in pre(S) \wedge n \in next(S) \quad i=0 \\ 1, & p \in pre(S) \otimes n \in next(S) \quad i=1 \\ 0.5, & \text{e is the synonym of p, or t is the synonym of n} \quad i=2 \\ 0.25, & \text{p only has one same character with e, or n only has one same character with t} \quad i=3 \end{cases}$$

Where pre(S) is the set of the ambiguity *S*'s environment which is consist of former word; *next(S)* is the set of the ambiguity *S*'s environment which is consist of latter word; *p* is the former word of the current ambiguity, *n* is the latter word of the current ambiguity.

In the model the synonym is defined as:

Let s1 and s2 are both words. If the first three bits of s1's code in Tongyici Cilin are same with the first three bits of s2's code in Tongyici Cilin, s1 is the synonym of s2, or s2 is the synonym of s1.

2.2 Unknown Words Recognition

In the process of unknown words recognition, we consider not only the inner information of unknown words, but also the environment of unknown words.

(1) Related definition (productivity): Productivity is the weight which measures the single character's location in the whole word.

If Ai is a single character, ti is the tag of Ai's location, let ti∈{B, M, S}, $P_{Ai}(t_i)$ is the productivity of the single character Ai in the location ti, which we can write as follows:

$$P_{A_i}(t_i) = \frac{count(A_i, t_i)}{\sum_{t_i \in T} count(A_i, t_i)}$$

(2)The inner information of unknown words mainly refer to the frequent of each character as word's begin, middle and end, as show in Table 1.

Word	Tag	Freq
A1	B/M/E	447/26/3
A2	B/M/E	2/0/0
A3	B/M/E	979/76/206
. . .	. . .	. . .

Table1 inner information of unknown words[3]

[2] HIT IR-Lab Tongyici Cilin (Extended)

[3] A1, A2, A3 represent the single character of Chinese. B, M, E represent respectively current character as the word's head, middle and end.

In the process of abstracting the exterior information, we have analyzed the tagged corpus and found that feature words have an important effect on the unknown words recognition, such as: predicate, post, specific behavior verb, etc.
For example:
Post: chairman, prime minister, etc.
Job：reporter, singer, writer, etc.
Appellation：comrade, sir, miss, etc.
Specific behavior verb：say, think, nominate, investigate, etc.

The process of unknown words recognition: "A1 A2 A3A4 A5A6….." is the disambiguation results, if single character of A2 has $P_{A2}(B)>0.35$ and $P_{A2}(M) >0.35$ or $P_{A2}(E) >0.35$, A1 A2 have the possibility to be an unknown words. After that, we filter it using the exterior information in order to improve R_{oov}.

3 Performance and analysis

The performance of our system in the SIGHAN bakeoff 2007 is presented in table 2.

OPEN	R	P	F	P_{OOV}	R_{IV}
NCC	94.5	92.6	93.5	71.6	96.9

Table 2 NCC test in SIGHAN bakeoff 2007 (%)

Our system has better performance in terms of Riv measure which attributed to the module of disambiguation. However, because the unsuitable threshold choice leads lots words combined incorrectly, the Roov measure is lower.

4 Conclusions

In this paper we use a cascade model to finish WS task and the system achieves a good performance on R_{iv} measure. It indicates that this method is feasible and effective. However, the shortcoming of the system is that the method of unknown words recognition hasn't got ideal performance which will be our future research focus.

References

Maosong Sun, Jiayan Zou, etc. 1999 .*The Role of High Frequent Maximal Crossing Ambiguities in Chinese Word Segmentation*. Journal of Chinese of Information Processing, 13(1):27-37

Kaiying Liu. 2000. *Automatic Chinese Word Segmentation and POS Tagging*. Business Publishing House. Beijing.

Luo Zhiyong, Song Rou. 2006. *Disambiguation in a Modern Chinese General-Purpose Word Segmentation System*. Journal of Computer Research and Development, 6:1122-1128.

Huang Changning, Zhao Hai. 2006. *Character-Based Tagging: A New Method for Chinese Word Segmentation*. Frontiers of Chinese Information Processing: 53-63.

Linxin, NetEase. 2006. *Automatic Chinese Word Segmentation.* Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney: 193–196.

Wu Liu, Heng Li. 2006. *France Telecom R&D Beijing Word Segmenter for Sighan Ba*keoff2006. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney: 193–196.