NOKIA Research Center Beijing Chinese Word Segmentation System for the SIGHAN Bakeoff 2007

Jiang LI[1, 2], Rile HU[1], Guohua ZHANG[1], Yuezhong TANG[1],
Zhanjiang SONG[1] and Xia WANG[1]
NOKIA Research Center, Beijing[1]
Beijing University of Posts and Telecommunications[2]
{ext-jiang.1.li, ext-rile.hu,ext-guohua.zhang,yuezhong.tang,
zhanjiang.song,xia.s.wang}@nokia.com

Abstract

This paper presents the Chinese word segmentation system developed by NOKIA Research Center (NRC), which was evaluated in the Fourth International Chinese Language Processing Bakeoff and the First CIPS Chinese Language Processing Evaluation organized by SIGHAN. In our system, a preprocessing module was used to discover the out-of-vocabulary words which occur repeatedly in the text, then an improved n-gram model was used for segmentation and some post processing strategies are adopted in system to recognize the organization names and new words. We took part in three tracks, which are called the open and closed track on corpora State Language Commission of P.R.C. (NCC), and closed track on corpora Shanxi University (SXU). Our system achieved good performance, especially in the open track on NCC, our system ranks 1st among 11 systems.

1 Introduction

Chinese word segmentation is an essential and core technology in Chinese language processing, and generally it is the first stage for later processing, such as machine translation, text summarization, information retrieval and etc. The topic of Chinese word segmentation has been researched for many years. Many approaches have been developed to solve the problems under this topic. Among these approaches, statistical approaches are most widely used.

Our system based on a pragmatic approach, integrating a lot of features and information, the framework is similar to (Jianfeng Gao, 2005). In our system, the model is simplified to n-gram architecture. First, all the possible paths of segmentation will be considered and each of the candidate word will be categorized into a certain type. Second, each word will be given a value; each type has different computational strategy and is processed in different ways. At last, all the possible paths of segmentation are calculated and the best path is selected as the final result.

N-gram language model is a generative model, and it could express the correlation of the context word very well. But it is powerless to detect the out-of-vocabulary word (OOV). In the post-processing module, we detect the OOV through some Chinese character information instead of the word information. In addition, to deal with the long organization names in NCC corpus, a module for combining organization name is adopted.

The remainder of this paper is organized as follow: section 2 describes our system in detail; section 3 presents the experiment results and analysis; in last section we give our conclusions and future research directions.

2 System Description

The basic architecture of our system is shown in figure 1, and the detailed description of each module is provided in the following subsections.

2.1 Framework

The input of the system is text to be segmented. First, the system scans the text and finds out the character strings appear many times but not lexicon words. These strings are called recurring-OOV. Second, all the candidate words are categorized into different types and the optimal path is calculated by Viterbi algorithm. Finally, some post-processing strategies are used to modify the results: NW detection is used to merge two single characters as a new word, and organization combination is provided to combine some words as an organization name.

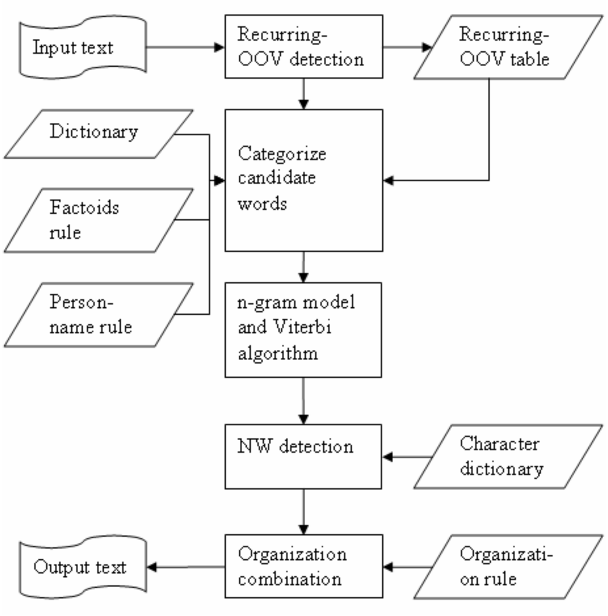

Figure 1: Framework of system

2.2 Recurring-OOV Detection

We found that there are some words which appear many times in different position of the context. For example, a verb "说给" appears 2 times in a sentence, "这是… **说给**大臣、太监听的，**说给**普天下劳苦大众的部分…", and a person name "**杨宇霆**" appears in different sentences, "**杨宇霆**感到身边四术士神机妙算…" "…**杨宇霆**向常荫槐使了个眼色…". These words are defined as Recurring-OOV.

Therefore, without any prior knowledge, the Chinese text is scanned; the sentence of the text is compared with itself and compared with others which were close to it for finding out the repeated strings. All these repeated character strings were saved in list. Not all of them are considered as the candidates of OOV word. Only 2-character or 3-character repeated strings are considered as the candidates of OOV words. And some simple rules are used to avoid some wrong classification. For example, if there is a repeated string contains character "的", which is a high frequent one-character word, this repeated string is not considered as a recurring OOV.

A value (probability) will be given to each Recurring-OOV. Two factors will be considered in the value evaluation：

1. The repeating times of the Recurring-OOV in the testing corpus. The more it repeats, the bigger the value will be.
2. Character-based statistical information and some other information are also considered to calculate the probability of this string to be a word. The computing method is described in Section 2.5, NW Detection.

2.3 Word Categorization

In our system, Chinese words are categorized into a set of types as follows:

1. Lexicon Words (LW). The words in this type can be found in the library we get from the training corpus.
2. Factoids (FT). This type includes the English letter, Arabic numerals and etc.
3. Named Entity (NE). This type includes person name and location name. Being different from the Gao's system, the organization detection is a post processing in our system.
4. Recurring-OOV. This type is described in the section 2.2.
5. Others.

2.4 N-gram Model

Each candidate word W_i is relegated to a type T_j and assigned a value $P(W_i \mid T_j)$. The transfer probability between word-types is also assigned a value $P(T_{j-1} \mid T_j)$. We assume that $W = W_1 W_2 \ldots W_n$ is a word sequence, the segmentation module's task is to find out the most likely word sequence among all possible paths:

$$W^* = \arg\max \prod P(W_i \mid T_j) P(T_{j-1} \mid T_j) \qquad (1)$$

Viterbi algorithm is used to search the optimized path described in Equation (1).

2.5 NW Detection

This module is used to detect the New Words, which “refer to OOV words that are neither recognized as named entities or factoids” (Jianfeng Gao, 2005). And in particular, we only consider the 2-character words for the reason that 2-character words are the most common in Chinese language.

We identify the new words through some features of Chinese character information:

1. The probability of a character occurring in the left/right position.

In fact, most Chinese characters have their favorite position. For example, “这” almost occurs in the left, “径” almost occurs in the right and “的” always compose a single word by itself. So the string “这x” is much more possible to be a new word than “x这”，and string “的x” is not likely to be a word.

2. The similarity of different characters.

If two characters often occur in the same position with the same character to form a word, it is considered that the two characters are similar, or there is a short distance between them. For example, the character “这” is very similar to “那” in respect that they are almost in the left position with some same characters, such as “里”, “么”, to construct the word “这里”,”那里”,”这么”,”那么”. So if we know the “这边” is a word, we can speculate the string “那边” is also a word.

The strict mathematical formula which used to describe the similarity of characters is reported in (Rile Hu, 2006).

2.6 Organization combination

The organization name is recognized as a long word in the NCC corpus, but during the n-gram processing, these long words will be segmented into several shorter words. In our system, the organization names are combined in this module. First, a list of suffix-words of organization name, such as “公司” “集团”, is selected from the training set. Second, the string that has been segmented in previous module is searched to find out the suffix-word, which is considered as a candidate of organization name. At last, we estimate the possibility of the candidate string and judge it is an organization name or not.

3 Evaluation Results

3.1 Results

We took part in three segmentation tasks in Bakeoff-2007, which are named as the open and closed track on corpora State Language Commission of P.R.C. (NCC), and closed track on corpora Shanxi University (SXU).

Precision (P), Recall (R) and F-measure (F) are adopted to measure the performance of word segmentation system. In addition, OOV Recall (R_{OOV}), OOV Precision (P_{OOV}) and OOV F-measure (F_{OOV}) are very important indicators which reflect the system’s ability to deal with the OOV words.

The results of our system in three tasks are shown in Table 1.

Table 1: Test set results on NCC, SXU

Corpus	NCC-O	NCC-C	SXU-C
R	0.9735	0.9417	0.9558
P	0.9779	0.9272	0.9442
F	0.9757	0.9344	0.95
R_{OOV}	0.8893	0.4001	0.5176
P_{OOV}	0.8867	0.6454	0.6966
F_{OOV}	0.888	0.494	0.5939
R_{IV}	0.9777	0.9687	0.9794
P_{IV}	0.9824	0.9356	0.9539
F_{IV}	0.98	0.9518	0.9665

3.2 NCC Open Track

For the open track of NCC, an external corpus is used for training and the size of training set is about 54M. In addition, there are some special dictionary were added to identify some special words. For example, an idiom dictionary is used to find the idioms and a personal-name dictionary is used to identify the common Chinese names.

3.3 Error Analysis

Apart from ranking 1st in NCC open test, our system got not so good results in NCC close test and SXU close test.

The comparison between our system results and best results in bakeoff-2007 are shown in table 2.

Table 2: The comparison between our system results and best results

Type	F-Measures	
	Bakeoff-2007	Our system
NCC-O	0.9757	0.9757
NCC-C	0.9405	0.9344
SXU-C	0.9623	0.95

Table 3: The comparison between our system results and best Top 3 results in OOV identification

		R_{OOV}	P_{OOV}	F_{OOV}
NCC-C	Our	0.4001	0.6454	0.494
	1st	0.6179	0.5984	0.608
	2nd	0.4502	0.6196	0.5215
	3rd	0.6158	0.5542	0.5834
SXU-C	Our	0.5176	0.6966	0.5939
	1st	0.7429	0.7159	0.7292
	2nd	0.6454	0.7022	0.6726
	3rd	0.6626	0.6639	0.6632

In table 3, 1st, 2nd and 3rd are the best Top 3 systems in the test. It shows that in the close track in NCC, the OOV Precision of our system is the best, but the OOV Recall is the worst in all the four system. Similarly, in the close track in SXU, the OOV Precision is very close to the best one, and the OOV Recall is the worst. It means that our system is too cautious in identifying the OOV words.

Our system was carefully tuned on NCC training set. The NCC training set contains articles from many domains; the OOV words can not be easily detected. Therefore, in parameter tuning, we raise the threshold of OOV. This strategy increases the precision of the OOV detection, but decreases the recall of this. And we also use some simple rules to filter the OOV candidates. These rules can easily pick out the wrongly detected OOVs, but at the same time, they remove some correct candidates by mistake.

The performance of our system is good in NCC close test but not so good in SXU close test. This means that our strategies for OOV detection is too cautious for SXU close test.

4 Conclusion and Future Work

In this paper, a detailed description on a Chinese word segmentation system is presented. N-gram model is adopted as the language model, and some preprocessing and post processing methods are integrated as a unified framework. The evaluation results show the efficiency of our approaches.

In future research, we will continue to enhance our system with other new techniques, especially we will focus on improving the recall of OOV words.

References

Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2005. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach. Computational Linguistics, Vol.31(4): 531-574.

Hai Zhao, Chang-Ning Huang and Mu Li. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field. Proceedings of the fifth SIGHAN Workshop on Chinese Language Processing, 162-165. Sydney, Australia.

Adwait Ratnaparkni. 1996. A Maximum Entropy Part-of-speech Tagger. In Proceedings of the Empirical Method in Natural Language Processing Conference, 133-142. University of Pennsylvania.

JK Low, HT Ng, W Guo. 2005. A Maximum Entropy Approach to Chinese Word Segmentation. Proceedings of the fourth SIGHAN Workshop on Chinese Language Processing. Jeju Island, Korea.

Manning, Christopher D. and Hinrich Schutze. 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, London, England.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. International Journal of Computational Linguistics and Chinese Language Processing, 8(1).

Rile Hu, Chengqing Zong, and Bo Xu. An Approach to Automatic Acquisition of Translation Templates Based on Phrase Structure Extraction and Alignment. *IEEE Transaction on Speech and Audio Processing.* Vol. 14, No.5, September 2006.