Description of the HKU Chinese Word Segmentation System for Sighan Bakeoff 2005

Guohong Fu
Department of Linguistics
The University of Hong Kong
Pokfulam Road, Hong Kong
ghfu@hkucc.hku.hk

Kang-Kwong Luke
Department of Linguistics
The University of Hong Kong
Pokfulam Road, Hong Kong
kkluke@hkusua.hku.hk

Percy Ping-Wai WONG
Department of Linguistics
The University of Hong Kong
Pokfulam Road, Hong Kong
wongpw@hkusua.hku.hk

Abstract

In this paper, we describe in brief our system for the Second International Chinese Word Segmentation Bakeoff sponsored by the ACL-SIGHAN. We participated in all tracks at the bakeoff. The evaluation results show our system can achieve an F measure of 0.940-0.967 for different testing corpora.

1 Introduction

Word segmentation is very important for Chinese text processing, which is aiming at recognizing the implicit word boundaries in plain Chinese text. Over the past decades, great progress has been made with Chinese word segmentation technology. However, two difficulties still face us while developing a practical segmentation system for large open applications, i.e. the resolution of ambiguous segmentation and the identification of unknown or out-of-vocabulary (OOV) words.

In order to resolve the above two problems, we developed a purely statistical Chinese word segmentation system using a two-stage strategy. We participated in eight tracks at the Second International Chinese Word Segmentation Bakeoff sponsored by the ACL-SIGHAN, and tested our system on different testing corpora. The scored results show that our system is effective for most of ambiguous segmentation and unknown words in Chinese text. In this paper, we make a summary of this work and give some analysis on the results.

The rest of this paper is organized as follows: First in Section 2, we describe in brief a two-stage strategy for Chinese word segmentation. Then in Section 3, we give details about the settings or configuration of our system for different testing tracks, particularly the training data and the dictionaries used in our system. Finally, we report the results of our system at this bakeoff in Section 4, and give our conclusions on this work in Section 5.

2 Overview of the System

In practice, our system works in two major steps as follows:

The first step is a process of known word segmentation, which aims to segment an input sequence of Chinese characters into a sequence of known words that are listed in the system dictionary. In our current system, we apply a known word bigram model to perform this task (Fu and Luke, 2003; Fu and Luke, 2004; Fu and Luke, 2005).

Actually, known word segmentation is a process of disambiguation. Given a Chinese character string $C = c_1 c_2 \cdots c_n$, there are usually multiple possible segmentations of known words $W = w_1 w_2 \cdots w_m$ according to the system dictionary. The task of known word segmentation is to find a proper segmentation $\hat{W} = w_1 w_2 \cdots w_m$ that maximizes the probability $\prod_{i=1}^{m} P(w_i \mid w_{i-1})$, i.e.

$$\hat{W} = \arg\max_{W} P(W \mid C) \approx \arg\max_{W} \prod_{i=1}^{m} P(w_i \mid w_{i-1}) \quad (1)$$

The second step is actually a tagging task on the sequence of known words acquired in the first step, which intends to detect unknown

words or out-of-vocabulary (OOV) words in the input. In this process, each known word yielded in the first step will be further assigned a proper tag that indicates whether the known word is an independent segmented word by itself or a beginning/middle/ending component of an OOV word (Fu and Luke, 2004). In order to improve our system, part-of-speech information is also introduced in some tracks such as the PKU open test and the AS open test. Furthermore, a lexicalized HMM tagger is developed to perform this task (Fu and Luke, 2004).

Given a sequence of known words $W = w_1 w_2 \cdots w_n$, the lexicalized HMM tagger attempt to find an appropriate sequence of tags $\hat{T} = t_1 t_2 \cdots t_n$ that maximizes the conditional probability $P(T \mid W)$, namely

$$\begin{aligned} \hat{T} &= \arg\max_T P(T \mid W) \\ &\approx \arg\max_T \prod_{i=1}^{n} P(w_i \mid w_{i-1}, t_i) P(t_i \mid w_{i-1}, t_{i-1}) \end{aligned} \quad (2)$$

3 Settings for Different Tracks

Table 1. Training corpora for different tracks

Table 1 presents the corpora used to train our system for different tracks. In the Academia Sinica (AS) open test and the Peking University (PKU) open test, our system is trained respectively using the *Sinica Corpus* (3.0) and *the PFR Corpus*. In all other tests, including all closed tests, City University of Hong Kong (CityU) open test and Microsoft Research (MSR) open test, we trained our system using the relevant training corpora provided for the bakeoff.

Track	Training Corpus	Word counts
AS-O	*The Sinica Corpus 3.0*	5,692,150
AS-C	AS corpus for Bakeoff 2005	5,449,698
CityU-O	CityU corpus for Bakeoff 2005	1,455,629
CityU-C	CityU corpus for Bakeoff 2005	1,455,629
MSR-O	MSR corpus for Bakeoff 2005	2,368,391
MSR-C	MSR corpus for Bakeoff 2005	2,368,391
PKU-O	*The PFR Corpus*	7,286,960
PKU-C	PKU corpus for Bakeoff 2005	1,109,947

Table 2 shows all the dictionaries used in our system for different tracks.

In the closed test, the system dictionaries are derived automatically from the relevant training corpora for this bakeoff by using the following three criteria: (1) Each character in the training corpus is taken as an independent entry and collected into the relevant system dictionary. (2) A standard Chinese word in the training corpus will enter to the relevant dictionary if it has four or less Chinese characters within it, and at the same time, its counts of occurrence in the corpus is observed to be larger than a threshold. In our current system, the threshold is set to 10 for the AS closed test and 5 for other closed tests. (3) For non-standard Chinese words such as numeral expressions, English words and punctuations, if they consist of multiple characters, they will be not included in the system dictionary.

As for the open test, some other dictionaries are applied. As can be seen from Table 2, *the CKIP Lexicon and Chinese Grammar* is used in both AS and CityU open test, and the *Grammatical Knowledge-base of Contemporary Chinese* developed by the Peking University is utilized in both PKU and MSR open test.

Track	System dictionary	# of entries
AS-O	*The CKIP Lexicon and Chinese Grammar*	84K
AS-C	Automatically extracted from AS corpus for Bakeoff 2005	30K
CityU-O	*The CKIP Lexicon and Chinese Grammar* (without part-of-speech)	84K
CityU-C	Automatically extracted from CityU corpus for Bakeoff 2005	22K
MSR-O	*The Grammatical Knowledge-base of Contemporary Chinese* (without part-of-speech)	65K
MSR-C	Automatically extracted from MSR corpus for Bakeoff 2005	17K
PKU-O	*The Grammatical Knowledge-base of Contemporary Chinese*	65K
PKU-C	Automatically extracted from PKU corpus for Bakeoff 2005	17K

Table 2. System dictionaries for different tracks

It should be noted that part-of-speech information is also utilized in the AS open test and the PKU open test, because part-of-speech information proved to be informative in identifying OOV words in Chinese text (Fu and Luke, 2004). Therefore, the training corpora for the two tests are tagged with part-of-speech, and entries in the relevant dictionaries are defined with their potential part-of-speech categories.

4 The Scored Results

In Bakeoff 2005, six measures are employed to score the performance of a word segmentation system, namely recall (*R*), precision (*P*), the evenly-weighted F-measure (*F*), out-of-vocabulary (OOV) rate for the test corpus, recall with respect to OOV words (R_{OOV}) or in-vocabulary words (R_{iv}).

In order to achieve a consistent evaluation of our system in both the closed test and the open test, OOV is defined in this paper as the set of words in the test corpus but not occurring in both the training corpus and the system dictionary. Furthermore, the additional two rates, i.e. OOV-C and OOV-D are used to denote the out-of-vocabulary rate with respect to the training corpus and the out-of-vocabulary rate with respect to the system dictionary, respectively. At the same time, the precision with regard to in-vocabulary words (P_{iv}) and OOV words (P_{OOV}) are also computed in this paper to give a more complete evalution of our system in unknown word identification.

Track	OOV-C	OOV-D	OOV
AS-O	0.043	0.096	0.039
AS-C	0.043	0.096	0.043
CityU-O	0.074	0.140	0.049
CityU-C	0.074	0.127	0.074
MSR-O	0.026	0.076	0.023
MSR-C	0.026	0.087	0.026
PKU-O	0.038	0.070	0.033
PKU-C	0.058	0.091	0.058

Table 3. OOV rates for different tracks

Track	F	R	P	R_{iv}	P_{iv}	R_{OOV}	P_{OOV}
AS-O	0.946	0.955	0.938	0.972	0.947	0.532	0.660
AS-C	0.940	0.947	0.934	0.966	0.949	0.523	0.566
CityU-O	0.941	0.944	0.938	0.962	0.956	0.592	0.599
CityU-C	0.939	0.944	0.933	0.969	0.952	0.626	0.677
MSR-O	0.967	0.969	0.966	0.978	0.977	0.586	0.537
MSR-C	0.962	0.962	0.962	0.972	0.977	0.592	0.499
PKU-O	0.962	0.959	0.965	0.963	0.970	0.835	0.816
PKU-C	0.944	0.943	0.944	0.961	0.958	0.656	0.700

Table 4. Scores for different tracks

The OOV rates and scores of our system are summarized respectively in Table 3 and Table 4. The results show that our system can achieve a F-measure of 0.940-0.967 for different testing corpora while the relevant OOV rates are from 0.023 to 0.074.

Although our system has achieved a promising performance, there is still much to be done to improve it. Firstly, our system is purely statistics-based, it cannot yield correct segmentations for all non-standard words (NSWs) such as numeral expressions and English strings in Chinese text. Secondly, known word segmentation and unknown word identification are taken as two independent stages in our system. This strategy is obviously simple and more easily applicable (Fu and Luke, 2003). Although the known word bigram model can partly resolve this problem, it is not always effective for some complicated strings that contains a mixture of ambiguities and unknown words, such as "31 日夜" and the fragment "中行长葛" in the sentence "中行长葛支行注重健身".

5 Conclusions

This paper presents a two-stage statistical word segmentation system for Chinese. We participated in all testing tracks at the second Sighan bakeoff. The scored results show that our system can achieve a F-measure of 0.940-0.967 as a whole for different corpora. This indicates that the proposed system is effective for most ambiguous segmentations and unknown words in Chinese test. For future work, we hope to improve our system by incorporating some pattern rules to handle complicated ambiguous fragments and non-standard words in Chinese text.

References

Guohong Fu, and Kang-Kwong Luke. 2003. A two-stage statistical word segmentation system for Chinese. In: *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan, 156-159.

Guohong Fu, and Kang-Kwong Luke. 2004. Chinese unknown word identification as known word tagging. In: *Proceedings of the Third IEEE International Conference on Machine Learning and Cybernetics (ICMLC 2004)*, Shanghai, China, 2612-2617.

Guohong Fu, and Kang-Kwong Luke. 2005. Chinese unknown word identification using class-based LM. *Lecture Notes in Computer Science (IJCNLP 2004)*, 3248: 704-713.