Nanjing Normal University Segmenter for the Fourth SIGHAN Bakeoff

Xiaohe CHEN, Bin LI, Junzhi LU, Hongdong NIAN, Xuri TANG
Nanjing Normal University,
122, Ninghai Road, Nanjing, P. R. China, 210097
chenxiaohe5209@msn.com,gothere@126.com,
lujunzhi@gmail.com,nianhong-dong@hotmail.com,
tangxuriyz@hotmail.com

Abstract

This paper expounds a Chinese word segmentation system built for the Fourth SIGHAN Bakeoff. The system participates in six tracks, namely the CityU Closed, CKIP Closed, CTB Closed, CTB Open, SXU Closed and SXU Open tracks. The model of Conditional Random Field is used as a basic approach in the system, with attention focused on the construction of feature templates and Chinese character categorization. The system is also augmented with some post-processing approaches such as the Extended Word String, model integration and others. The system performs fairly well on the 5 tracks of the Bakeoff.

1 Introduction

The Nanjing Normal University (NJNU) team participated in CityU Closed, CKIP Closed, CTB Closed, CTB Open, SXU Closed, SXU Open tracks in the WS bakeoff. The system employed in the Bakeoff is based mainly on the model of CRF, optimized with some pre-processing and post-processing methods. The team has focused its attention on the construction of feature templates, Chinese character categorization, the use of Extended Word String and the integration of different segmentation models in the hope of achieving better performance in both IVs（In Vocabulary words） and OOVs (Out Of Vocabulary words). Due to time limitations, some of these methods are still not fully explored. However, the Bakeoff results show that the performance of the overall system is fairly satisfactory.

The paper is organized as follows: section 2 gives a brief description of the system; section 3 and 4 are devoted to the discussion of the results of closed test and open test; a conclusion is given to comment on the overall performance of the system.

2 System Description

Conditonal Ramdom Field (CRF) has been widely used by participants in the basic tasks of NLP since Peng(2004). In both SIGHAN 2005 and 2006 Bakeoffs CRF-based segmenters prove to have a better performance over other models. We have also chosen CRF as the basic model for the task of segmentation and uses the package CRF++ developed by Taku Kudo[1]. Some post-processing optimizations are also employed to improve the overall segmentation performance. The general description of the system is illustrated in Figure 1. The basic segmenter and post-processing are explained in the next two sections.

2.1 Basic Segmenter

As in many other segmentation models, our system also treats word segmentation as a task of classification problem. During the experiment of the model, two aspects are taken into consideration, namely tag set and feature template. The 6-tag (Table 1) set proposed in Zhao(2006) is employed to mark various character position status in a Chinese word. The feature template (Table 2) consid-

[1] Package CRF++, version 0.49, available at http://crfpp.sourceforge.net.

ers three templates of character features and three templates of character type features. The introduction of character type (Table 3) is based on the observation that many segmentation errors are caused by different segmentation standards among different corpora, especially between Traditional Chinese corpora and Simplified Chinese Corpora.

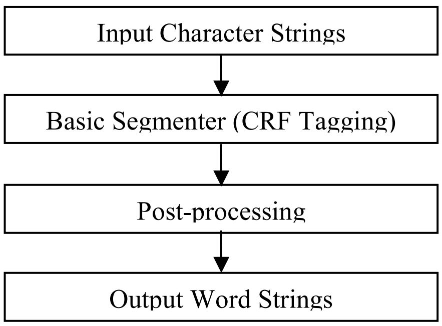

Figure 1: Flow Chat

Status	Tag
begin	B
2nd	B2
3rd	B3
middle	M
end	E
single	S

Table 1:6-tag Set

Type	Feature	Function
Char Unigram	C_n, n=-2, -1, 0, 1, 2	Character in position n to the current character
Char Bigram	C_nC_{n+1}, n=-1,0	Previous(next) character and current character
Char Jump	C_{-1} C_1	Previous character and next character
CharType Unigram	T_n, n=-1, 0, 1	Type of previous (current, next) character
CharType Bigram	T_nT_{n+1}, n=-1,0	Type of previous character and next character
CharType Jump	T_{-1} T_1	Type of previous character and next character

Table 2: Feature Templates in Close Test

Character Type	Example
Chinese Character	我 人
Serial Number	1.⑴ ①㈠
Roman Number	Ⅰ Ⅱ ⅷ
Aribic Number	12１２
Chinese Number	零〇百壹
Ganzhi	甲乙子丑
Foreign Character	Ａ Δ は
National Pronunciation Letters	ㄅㄉ ˇ
Sentence Punctuation	；！。？
Hard Punctuation	\t\r\n
Punctuation	：…-¨"
Dun	、 `
Dot1	‥
Dot2	. ·
Di	第
At	@
Other Character	⊙∽

Table 3:Character Type

2.2 Post-Processing

Two methods are used in post-processing to optimize the results obtained from basic segmenter. The first is the binding of digits and English Characters. The second is the use of extended word string to solve segmentation ambiguity.

2.2.1 Binding Digits and Roman Letters

Digits (ranging from "0" to "9") are always bound as a word in Chinese corpora, while roman letters are treated differently in different corpora, some adding a full-length blank between the letters, some not. The system employs rule-based approach to bind both digits and roman letters. We also submitted two segmentation results for the Bakeoff, please refer to section 3.2 for discussion of these results.

2.2.2 Extended Word String (EWS) Approach

The CRF model performs well in segmenting IV word strings in general, but not in all contexts. Our system thus uses a memory based method, which is named as Extended Word String approach, to prevent CRF from making such error. All the Chinese word strings, which are of character length from 2 to 10 and appear more than two times, are stored in a hash table, together with information of their segmentation forms. An example of EWS is given in Table 5. If the same character string appears in the test data, the system can easily re-segment them by querying the hash table. If the query finds that the character string has only one segmentation form and checking shows that the string has no overlapping ambiguity with its left or right word, the segmentation of the string is then modified according to the stored segmentation type. Our experiment shows that the approach can promote the F-measure by 0.2% to 1% on different tracks.

EWS	Seg Form	Freq
就我们	/就/我们/	4

Table 5: Example of EWS

3 Evaluation Results on Closed Test

3.1 CKIP Closed Test

In CKIP Closed Test, another kind of post processing is used for OOVs. Examination on the output from basic segmenter shows that some OOVs identified by CRFs are not OOV errors, but IV errors. Sometimes it can not always segment the same OOV correctly in different context. For example, the person name "陳子江" appears three times in the test, but it is only correctly detected twice, and for once it is wrongly detected. Our approach is to re-segment the OOVs string (with its left and right word) twice. Firstly the string is segmented using the training data wordlist, followed by a second segmentation using the OOV wordlist recognized by the Basic Segmenter. The result with the minimum number of words is accepted.

```
Example:
Basic Seg Output：/的/陳子/江本/
OOV Adjusting：   /的/陳子江/本/
Basic Seg Output：/血永/不融/和/
OOV Adjusting：   /血/永不/融和/
```

With the OOV Adjusting Approach mentioned above, we got the third place in the track (Table 6). But when we use it on other corpora, the method does not promote the performance. Rather, it lowers the performance score. The reason is still not clear.

System（rank）	F	F_{oov}	F_{iv}
Best(1/21)	0.9510	0.7698	0.9667
Njnu(3/21)	0.9454	0.7475	0.9637

Table 6: CityU Closed Test

3.2 CKIP and CTB Closed Test

In CKIP Closed Test, only the basic segmenter introduced in section 2 is used. Two segmentation results, namely ***a*** and ***b*** (Table 7 and 8) are submitted for the Bakeoff. Result ***a*** binds the roman letters as a word, while result ***b*** does not. The scores of the two results show that the approach is not stable in terms of score. We suggest that corpora submitted for evaluation purposes should pay more attention to non-Chinese word tagging and comply with the request of Bakeoff organizers.

System（rank）	F	F_{oov}	F_{iv}
Best(1/19)	0.9470	0.7524	0.9623
Njnu a(6/19)	0.9378	0.6948	0.9580
Njnu b(9/19)	0.9204	0.6341	0.9452

Table 7: CKIP Closed Test

System（rank）	F	F_{oov}	F_{iv}
Best(1/26)	0.9589	0.7745	0.9697
Njnu a(9/26)	0.9498	0.7152	0.9645
Njnu b(7/26)	0.9499	0.7142	0.9647

Table 8: CTB Closed Test

3.3 SXU Closed Test

Four results (***a***, ***b***, ***c*** and ***d***) are submitted for this track (Table 9). Results ***a*** and ***b*** are dealt in the same way as described in section 3.2. Result ***c*** is obtained by incorporating results from a memory-based segmenter. The memory-based segmenter is mainly based on memory-based learning proposed by Daelemans(2005). We tested it on the training data with 90% as training data and 10% as testing data. The result shows that performance is improved. However, when the method is applied on the Bakeoff test data, the performance is lowered. The reason is not identified yet.

Result ***d*** was based on result ***c***. It incorporates OOV words recognized by the system introduced in (Li & Chen, 2007) in the post-processing stage. Based on suffix arrays, Chinese character strings with mutual information value above 8.0 are automatically extracted as words without any manual operation. We can see from table 9 that the F-measure of result ***d*** improved and F_{oov} of ***d*** got 2rd place in the test. And it is likely to get higher score if we combine it with result ***a***.

System（rank）	F	F_{oov}	F_{iv}
Best(1/29)	0.9623	0.7292	0.9752
Njnu a(9/29)	0.9539	0.6789	0.9702
Njnu b(10/29)	0.9538	0.6778	0.9701
Njnu c(15/29)	0.9526	0.6793	0.9688
Njnu d(14/29)	0.9532	0.6817	0.9694

Table 9: Sxu Closed Test

4 Evaluation Results on Open Test

4.1 Methods

More features and resources are used in open test, mainly applied in the modification of feature templates. Besides the features used in the close test, we add to feature templates more information about Chinese characters, such as the Chinese radicals ("扌口"), tones (5 tones), and another 6 Boolean values for each Chinese character. The 6 Boolean values indicate respectively whether the character is of Chinese surnames ("张王"), or of Chinese names ("琴林"), or of characters used for western person name translation ("尼克"), or of character used for English location name translation("纽约"), or of affixes ("老-","-者"), or of single character words ("了他"). The feature templates constructed in this way is given in Table 10.

Type	Feature	Function
Char Unigram	C_n, n=-1,0,1	The prevoius (current, next) character
Char Bigram	$C_n C_{n+1}$, n=-1,0	The previous(next) character and current character
Char Jump	$C_{-1} C_1$	The previous character and next character
CharType Unigram	T_0	The type of the current, next character
CharType Trigram	$T_{-1} T_0 T_1$	The type of the previous, current and next character
Char Information Unigram	T_0^n, n=1,…,6	The 6 information of the current, next character
Char Information Trigram	$T_{-1}^n T_0^n T_1^n$, n=1,…,6	The 6 information of the previous, current and next character

Table10: Feature Templates for Open Test

In the post-processing stage, we also add a Chinese idiom dictionary (about 27000 items) to help increase the OOV word recall.

4.2 Results

In SXU open test, we submitted 3 results (***a***, ***b*** and ***c***), but only ***a*** achieves the 4th rank in F-measure (Table 11). Features and resources added to the system turns out not to be of much use in the task, compared with our score on the closed test.

Result ***b***, ***c*** and all the results in CTB open test submitted have errors due to our pre-processing stage with CRF. Thus, the scores of them are very low, and some are even lower than our scores in closed test (see table 12).

System（rank）	F	F_{oov}	F_{iv}
Best(1/9)	0.9735	0.8109	0.9820
Njnu a(4/12)	0.9559	0.6925	0.9714

Table 11: SXU Open Test

System（rank）	F	F_{oov}	F_{iv}
Best(1/12)	0.9920	0.9654	0.9936
Njnu a(9/12)	0.9346	0.6341	0.9528

Table 12: CTB Open Test

5 Conclusions and Future Work

This is the first time that the NJNU team takes part in SIGHAN WS Bakeoff. In the construction of the system, we conducted experiments on the CRF-based segmenter with different feature templates. We also employs different post-processing approaches, including Extended Word String approach, digit and western roman letter combination, and OOV detection. An initial attempt is also made on the integration of different segmentation models. Time constraint has prevented the team from fuller exploration of the methods used in the system. Future efforts will be directed towards more complicated segmentation models, the examination of the function of different features in the task, the integration of different models, and more efficient utility of other relevant resources.

References

Bin Li, Xiaohe Chen. 2007. A Human-Computer Interaction Word Segmentation Method Adapting to Chinese Unknown Texts*, Journal of Chinese Information Processing*, 21(3):92-98.

Daelemans, W. and Van den Bosch. 2005. *Memory-Based Language Processing*. Cambridge University Press, Cambridge, UK.

Fuchun Peng, et al. 2004. Chinese Segmentation and New Word Detection Using Conditional Random Fields, *COLING2004*, 562-568, 23-27 August, Geneva, Switzerland.

Gina-Anne Levow. 2006. The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition, *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108-117, 22-23 July, Sydney, Australia.

Hai Zhao, et al. 2006. An Improved Chinese Word Segmentation System with Conditional Random Field, *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing,* 162-165, 22-23 July, Sydney, Australia.

Richard Sproat and Thomas Emerson. 2003. The First International Chinese Word Segmentation Bakeoff, *The Second SIGHAN Workshop on Chinese Language Procesing*, 133-143, Aspporo, Japan.

Thomas Emerson. 2005. The Second International Chinese Word Segmentation Bakeoff, *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 123-133, Jeju Island, Korea.