High OOV-Recall Chinese Word Segmenter

Xiaoming Xu, Muhua Zhu, Xiaoxu Fei, and Jingbo Zhu
School of
Information Science and Engineering
Northeastern University
{xuxm, zhumh, feixx}@ics.neu.edu.cn
zhujingbo@mail.neu.edu.cn

Abstract

For the competition of Chinese word segmentation held in the first CIPS-SIGHNA joint conference. We applied a subword-based word segmenter using CRFs and extended the segmenter with OOV words recognized by Accessor Variety. Moreover, we proposed several post-processing rules to improve the performance. Our system achieved promising OOV recall among all the participants.

1 Introduction

Chinese word segmentation is deemed to be a prerequisite for Chinese language processing. The competition in the first CIPS-SIGHAN joint conference put the task of Chinese word segmentation in a more challengeable setting, where training and test data are obtained from different domains. This setting is widely known as *domain adaptation*.

For domain adaptation, either a large-scale unlabeled target domain data or a small size of labeled target domain data is required to adapt a system built on source domain data to the target domain. In this word segmentation competition, unfortunately, only a small size of unlabeled target domain data is available. Thus we focus on handling out-of-vocabulary (OOV) words. For this purpose, our system is based on a combination of subword-based tagging method (Zhang et al., 2006) and accessor variety-based new word recognition method (Feng et al., 2004). In more detail, we adopted and extended subword-based method. Subword list is augmented with new-word list recognized by accessor variety method.

Feature Template	Description
a) $c_n(-2,-1,0,1,2)$	unigram of characters
b) $c_n c_{n+1}(-2,-1,0,1)$	bigram of characters
c) $c_{n-1}c_n c_{n+1}(-1,0,1)$	trigram of characters
d) $P_u(C_0)$	whether punctuation
e) $T(C_{-1})T(C_0)T(C_{+1})$	type of characters

Table 1: Basic Features for CRF-based Segmenter

We participated in the close track of the word segmentation competition, on all the four test datasets, in two of which our system is ranked at the 1st position with respect to the metric of OOV recall.

2 System Description

2.1 Subword-based Tagging with CRFs

The backbone of our system is a character-based segmenter with the application of Conditional Random Fields (CRFs) (Zhao and Kit, 2008). In detail, we apply a six-tag tagging scheme, as in (Zhao et al., 2006). That is , each Chinese character can be assigned to one of the tags in $\{B$, B_2, B_3, M, E, S $\}$. Refer to (Zhao et al., 2006) for detailed meaning of the tags. Table 1 shows basic feature templates used in our system, where feature templates a, b, d, e are also used in (Zhu et al., 2006) for SVM-based word segmentation.

In order to extend basic CRF-based segmenter, we first collect 2k most frequent words from training data. Hereafter, the list of such words is referred to as *subword list*. Moreover, single-character words [1], if they are not contained in the subword list, are also added. Such proce-

[1] By single-character word, we refer to words that consist solely of a Chinese character.

Feature Template	Description
f) in(str, subword-list)	is str in subword list
g) in(str, confident-word-list)	is str in confident-word list

Table 2: Subword Features for CRF-based Segmenter

dure for constructing a subword list is similar to the one used in (Zhang et al., 2006). To enhance the effect of subwords, we go one step further to build a list, named *confident-word list* here and below, which contains words that are not a portion of other words and are never segmented in the training data. In the competition, 400 most frequent words in the confident-word list are used. With subword list and confident-word list, both training and test data are segmented with forward maximum match method by using the union of subword list and confident-word list. Each segmentation unit (single-character or multi-character unit) in the segmentation results are regarded as "pseudo character" and thus can be represented with the basic features in Table 1 and two additional features as shown in Table 2. See the details of subword-based Chinese word segmentation in (Zhang et al., 2006)

2.2 OOV Recognition with Accessor Variety

Accessor variety (AV) (Feng et al., 2004) is a simple and effective unsupervised method for extraction of new Chinese words. Given a unsegmented text, each substring (candidate word) in the text can be assigned a value according to the following equation:

$$AV(s) = min\{L_{av}(s), R_{av}(s)\} \qquad (1)$$

where the left and right AV values, $L_{av}(s)$ and $R_{av}(s)$ are defined to be the number of distinct character types appearing on the left and right, respectively. Candidate words are sorted in the descending order of AV values and most highly ranked ones can be chosen as new words. In practical applications, heuristic filtering rules are generally needed (Feng et al., 2004). We re-implemented the AV method and filtering rules, as in (Feng et al., 2004). Moreover, we filter out candidate words that have AV values less than 3. Unfortunately, candidate word list generated this way still contains many noisy words (substrings that are not words). One possible reason is that unlabeled data (test data) used in the competition is extremely small in size. In order to refine the results derived from the AV method, we make use of the training data to filter the results from two different perspectives.

- Segment test data with the CRF-based segmenter described above. Then we collect (candidate) words that are in the CRF-based segmentation results, but not appear in the training data. Such words are called *CRF-OOV words* hereafter. We retain the intersection of CRF-OOV words and AV-based results as the set of candidate words to be processed by the following step.

- Any candidate word in the intersection of CRF-based and AV-based results will be filtered out if they satisfy one of the following conditions: 1) the candidate word is a part of some word in the training data; 2) the candidate word is formed by connection of consecutive words in the training data; 3) the candidate word contains position words, such as 上 (up), 下 (down), 左 (left), 右 (right), etc.

Moreover, we take all English words in test data as OOV words. A simple heuristic rule is defined for the purpose of English word recognition: an English word is a consecutive sequence of English characters and punctuations between two English characters (including these two characters).

We finally add all the OOV words into subword list and confident-word list.

3 Post-Processing Rules

In the results of subword-based word segmentation with CRFs, we found some errors could be corrected with heuristic rules. For this purpose, we propose following post-processing rules, for handling OOV and in-vocabulary (IV) words, respectively.

3.1 OOV Rules

3.1.1 Annotation-Standard Independent Rules

We assume the phenomena discussed in the following are general across all kinds of annotation

standards. Thus corresponding rules can be applied without considering annotation standards of training data.

- A punctuation tends to be a single-character word. If a punctation's previous character and next character are both Chinese characters, i.e. not punctuation, digit, or English character, we always regard the punctuation as a word.

- Consecutive and identical punctuations tend to be joined together as a word. For example, "—" represents a Chinese hyphen which consists of three "-", and "!!!" is used to show emphasizing. Inspired by this observations, we would like to unite consecutive and identical punctuations as a single word.

- When the character "·" appears in the training data, it is generally used as a connections symbol in a foreign person name, such as "圣·约翰 (Saint John)". Taking this observation into consideration, we always unite the character "·" and its previous and next segment units into a single word. A similar rule is designed to unite consecutive digits on the sides of the symbol ".", ex. "1.11".

- We notice that four consecutive characters which are in the pattern of *AABB* generally form a single word in Chinese, for example "平平淡淡 (dull)". Taking this observation into account, we always unite consecutive characters in the *AABB* into a single word.

3.1.2 Templates with Generalized Digits

Words containing digits generally belong to a open class, for example, the word "2012年 (AD 2012）" means a date. Thus CRF-based segmenter has difficulties in recognizing such words since they are frequently OOV words. To attack this challenge, we first generalize digits in the training data. In detail, we replaced consecutive digits with "*". For example, the word "2012年" will be transformed into "*年". Second, we collect word templates which consist of three consecutive words on condition that at least one of the words in a template contains the character "*" and that the template appears in the training data more than 4 times. For example, we can get a template like "*月(month) *日(day) 电(publish)". With such templates, we are able to correct errors, say "10月 17日 电" into "10月 17日 电".

3.2 IV Rules

We notice that long words have less ambiguity than short words in the sense of being words. For example, characters in "人才 济济 （full of talents)" always form a word in the training data, whereas "人才" have two plausible splitting forms, as "人才 (talent)" or "人 (people) 才 (only)". In our system, we collect words that have at least four characters and filter out words which belong to one of following cases: 1) the word is a part of other words; 2) the word consists solely of punctation and/or digit. For example, "唯物主义 (materialism)" and "一百二十 (120)" are discarded, since the former is a substring of the word "唯物主义者 (materialist)" and the latter is a word of digits. Finally we get a list containing about 6k words. If a character sequence in the test data is a member in the list, it is retained as a word in the final segmentation results.

Another group of IV rules concern character sequences that have unique splitting in the training data. For example, "女人们 (women)" is always split as "女人 (woman) 们 (s)". Hereafter, we refer to such character sequences as *unique-split-sequence (USS)*. In our system, we are concerned with UUSs which are composed of less than 5 words. In order to apply UUSs for post-processing, we first collect word sequence of variable length (word number) from training data. In detail, we collect word sequences of two words, three words, and four words. Second, word sequences that have more than one splitting cases in the training data are filtered out. Third, spaces between words are removed to form USSs. For example, the words "女人 (woman) 们 (s)" will form the USS "女人们 ". Finally, we search the test data for each USS. If the searching succeeds, the USS will be replaced with the corresponding word sequence.

4 Evaluation Results

We evaluated our Chinese word segmenter in the close track, in four domain: literature (Lit), com-

Domain	Basic			+OOV			+OOV+IV		
	R_{OV}	R_{IV}	F	R_{OV}	R_{IV}	F	R_{OV}	R_{IV}	F
Lit	.643	.946	.927	.652	.947	.929	.648	.952	.934
Com	.839	.961	.938	.850	.961	.941	.852	.965	.947
Med	.725	.938	.912	.754	.939	.917	.756	.944	.923
Fin	.761	.956	.932	.854	.958	.950	.871	.961	.955

Table 3: Effectiveness of post-processing rules

puter (Com), medicine (Med) and finance (Fin). The results are depicted in Table 4, where R, P and F refer to Recall, Precision, F measure respectively, and R_{OOV} and R_{IV} refer to recall of OOV and IV words respectively. Since OOV words are the obstacle for practical Chinese word segmenters to achieve high accuracy, we have special interest in the metric of OOV recall. We found that our system achieved high OOV recall [2]. Actually, OOV recall of our system in the domains of *computer* and *finance* are both ranked at the 1st position among all the participants. Compared with the systems ranked second in these two domains, our system achieved OOV recall $.853\ vs.\ .827$ and $.871\ vs.\ .857$ respectively.

We also examined the effectiveness of post-processing rules, as shown in Table 3, where *Basic* represents the performance achieved before post-processing, *+OOV* represents the results achieved after applying OOV post-processing rules, and *+OOV+IV* denotes the results achieved after using all the post-processing rules, including both OOV and IV rules. As the table shows, designed post-processing rules can improve both IV and OOV recall significantly.

Domain	R	P	F	R_{OOV}	R_{IV}
Lit	.931	.936	.934	.648	.952
Com	.948	.945	.947	.853	.965
Med	.924	.922	.923	.756	.944
Fin	.953	.956	.955	.871	.961

Table 4: Performance of our system in the competition

[2]For the test data from the domain of literature, we actually use combination of our system and forward maximum match, so we will omit the results on this test dataset in our discussion.

5 Conclusions and Future Work

We proposed an approach to refine new words recognized with the accessor variety method, and incorporated such words into a subword-based word segmenter. We found that such method could achieve high OOV recall. Moreover, we designed effective post-processing rules to further enhance the performance of our systems. Our system finally achieved satisfactory results in the competition.

Acknowledgments

This work was supported in part by the National Science Foundation of China (60873091).

References

Feng, Haodi, Kang Chen, Xiaotie Deng, and Weimin zhang. 2004. *Accessor Variety Criteriafor Chinese Word Extraction.* Computational Linguistics 2004, 30(1), pages 75-93.

Zhang, Ruiqiang, Genichiro Kikui, and Eiichiro Sumita. 2006. *Subword-based Tagging by Conditional Random Fileds for Chinese Word Segmentation.* In Proceedings of HLT-NAACL 2006, pages 193-196.

Zhao, Hai, Chang-Ning Huang, and Mu Li. 2006. *Improved Chinese Word Segmentation System with Conditional Random Field.* In Proceedings of SIGHAN-5 2006, pages 162-165.

Zhao, Hai and Chunyu Kit. 2008. *Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition.* In Proceedings of SIGHAN-6 2008, pages 106-111.

Zhu, Muhua, Yiling Wang, Zhenxing Wang, Huizhen Wang, and Jingbo Zhu. 2006. *Designing Special Post-Processing Rules for SVM-based Chinese Word Segmentation.* In Proceedigns of SIGHAN-5 2006, pages 217-220.