POC-NLW Template for Chinese Word Segmentation

Bo Chen
chb615@gmail.com
Weiran Xu
xuweiran@263.net
Tao Peng
ppttbupt@gmail.com
Jun Guo
guojun@bupt.edu.cn
Pattern Recognition and Intelligent System Lab
Beijing University of Posts and Telecommunications
Beijing 100876, P. R. China

Abstract

In this paper, a language tagging template named POC-NLW (position of a character within an n-length word) is presented. Based on this template, a two-stage statistical model for Chinese word segmentation is constructed. In this method, the basic word segmentation is based on n-gram language model, and a Hidden Markov tagger based on the POC-NLW template is used to implement the out-of-vocabulary (OOV) word identification. The system participated in the MSRA_Close and UPUC_Close word segmentation tracks at SIGHAN Bakeoff 2006. Results returned by this bakeoff are reported here.

1 Introduction

In Chinese word segmentation, there are two problems still remain, one is the resolution of ambiguity, and the other is the identification of so-called out-of-vocabulary (OOV) or unknown words. In order to resolve these two problems, a two-stage statistical word segmentation strategy is adopted in our system. The first stage is optional, and the whole segmentation can be accomplished in the second stage. In the first stage, the n-gram language model is employed to implement basic word segmentation including disambiguation. In the second stage, a language tagging template named POC-NLW (position of a character within an n-length word) is introduced to accomplish unknown word identification as template-based character tagging.

The remainder of this paper is organized as follows. In section 2 and section 3, a briefly description of the main methods adopted in our system is given. Results of our system at this bakeoff are reported in section 4. At last, conclusions are derived in section 5.

2 The Basic Word Segmentation Stage

In the first stage, the basic word segmentation is accomplished. The key issue in this stage is the ambiguity problem, which is mainly caused by the fact that a Chinese character can occur in different word internal positions in different words (Xue, 2003). A lot of machine learning techniques have been applied to resolve this problem, the n-gram language model is one of the most popular ones among them (Fu and Luke, 2003; Li et al., 2005). As such, we also employed n-gram model in this stage.

When a sentence is inputted, it is first segmented into a sequence of individual characters (e.g. ASCII strings, basic Chinese characters, punitions, numerals and so on), marked as $C_{1,n}$. According to the system's dictionary, several word sequences $W_{1,m}$ will be constructed as candidates. The function of the n-gram model is to find out the best word sequence W^* corresponds to $C_{1,n}$, which has the maximum integrated probability, i.e.,

$$W^* = \arg\max_{W_{1,m}} P(W_{1,m} \mid C_{1,n})$$

$$\cong \arg\max_{W_{1,m}} \prod_{i=1}^{m} P(W_i \mid W_{i-1}) \qquad \textit{for bigram}$$

$$\cong \arg\max_{W_{1,m}} \prod_{i=1}^{m} P(W_i \mid W_{i-1}, W_{i-2}) \quad \textit{for trigram}$$

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 177–180,
Sydney, July 2006.

The Maximum Likelihood method was used to estimate the word n-gram probabilities used in our model, and the linear interpolation method (Jelinek and Mercer, 1980) was applied to smooth these estimated probabilities.

3 The OOV Word Identification Stage

The n-gram method is based on the exiting grams in the model, so it is good at judging the connecting relationship among known words, but does not have the ability to deal with unknown words in substance. Therefore, another OOV word identification model is required.

OOV words are regarded as words that do not exist in a system's machine-readable dictionary, and a more detailed definition can be found in (Wu and Jiang, 2000). In general, Chinese word can be created through compounding or abbreviating of most of existing characters and words. Thus, the key to solve the OOV word identification lies on whether the new word creation mechanisms in Chinese language can be extracted. Therefore, a POC-NLW language tagging template is introduced to explore such information on the character-level within words.

3.1 The POC-NLW Template

Many character-level based works have been done for the Chinese word segmentation, including the LMR tagging methods (Xue, 2003; Nakagawa. 2004), the IWP mechanism (Wu and Jiang, 2000). Based on these previous works, this POC-NLW template was derived. Assume that the length of a word is the number of component characters in it, the template is consist of two component: L_{max} and a *Wl-Pn* tag set. L_{max} to denote the maximum length of a word expressed by the template; a *Wl-Pn* tag denotes that this tag is assigned to a character at the *n-th* position within a *l*-length word, $n = 1,2,\cdots,l$. Apparently, the size of this tag set is $(L_{\max}+1)\times L_{\max}/2$

For example, the Chinese word "人民" is tagged as:

人 W2P1, 民 W2P2

and "中国人" is tagged as:

中 W3P1, 国 W3P2, 人 W3P3

In the example, two words are tagged by the template respectively, and the Chinese character "人" has been assigned two different tags.

In a sense, the Chinese word creation mechanisms could be extracted through statistics of the tags for each character on a certain large corpus.

On the other hand, while a character sequence in a sentence is tagged by this template, the word boundaries are obvious. Meanwhile, the word segmentation is implemented.

In addition, in this template, known words and unknown words are both regarded as sequences of individual characters. Thus, the basic word segmentation process, the disambiguation process and the OOV word identification process can be accomplished in a unified process. Thereby, this model can also be used alone to implement the word segmentation task. This characteristic will make the word segmentation system much more efficient.

3.2 The HMM Tagger

Form the description of POC-NLW template, it can be found that the word segmentation could be implemented as POC-NLW tagging, which is similar to the so-called part-of-speech (POS) tagging problem. In POS tagging, Hidden Markov Model (HMM) was applied as one of the most significant methods, as described in detail in (Brants, 2000). The HMM method can achieve high accuracy in tagging with low processing costs, so it was adopted in our model.

According to the definition of POC-NLW template, the state set of HMM corresponds to the *Wl-Pn* tag set, and the symbol set is composed of all characters. However, the initial state probability matrix and the state transition probability matrix are not composed of all of the tags in the state set. To express more clearly, we define two subset of the state set:

- Begin Tag Set (BTS): this set is consisted of tag which can occur in the begging position in a word. Apparently, these tags must have the *Wl-P1* form.
- *End Tag Set (ETS):* correspond to BTS, tags in this set should occur in the end position, and their form should be like *Wl-Pl*.

Apparently, the size of BTS is L_{max} as well as of ETS. Thus, the initial state probability matrix corresponds to BTS instead of the whole state set. On the other hand, because of the word internal continuity, if the current tag *Wl-Pn* is not in ETS, than the next tag must be *Wl-P(n+1)*. In other words, the case in which the transition probability is need is that when the current tag is in ETS and the next tag belongs to BTS. So, the state transition matrix in our model corresponds to $ETS \times BTS$.

The probabilities used in HMM were defined similarly to those in POS tagging, and were estimated using the Maximum Likelihood method from the training corpus.

In the two-stage strategy, the output word sequence of the first stage is transferred into the second stage. The items in the sequence, including individual characters and words, which do not have a bigram or trigram relationship with the surrounding items, are picked out with its surrounding items to compose several sequences of items. These item sequences are processed by the HMM tagger to form new item sequences. At last, these processed items sequences are combined into the whole word sequence as the final output.

4 Results and Analysis

4.1 System

The system submitted at this bakeoff was a two-stage one, as describe at beginning of this paper. The model used in the first stage was trigram, and the L_{max} of the template used in the second stage was set to 7.

In addition to the tags defined in the template before, a special tag is introduced into our *Wl-Pn* tag set to indicate all those characters that occur after the L_{max}-*th* position in an extremely long (longer than L_{max}) word., formulized as WL_{max}-$P(L_{max}+1)$. And then, there are 28 basic tags (from *W1-P1* to *W7-P7*) and the special one *W7-P8*.

For instance, using the special tag, the word "中国共产党中央委员会" (form the MSRA Corpus) is tagged as:

中 *W7-P1* 国 *W7-P2* 共 *W7-P3* 产 *W7-P4* 党 *W7-P5* 中 *W7-P6* 央 *W7-P7* 委 *W7-P8* 员 *W7-P8* 会 *W7-P8*

4.2 Results at SIGHAN Bakeoff 2006

Our system participated in the MSRA_Close and UPUC_Close track at the SIGHAN Bakeoff 2006. The test results are as showed in Table 1.

Corpus	MSRA	UPUC
F-measure	0.951	0.918
Recall	0.956	0.932
Precision	0.947	0.904
IV Recall	0.972	0.969
OOV Recall	0.493	0.546
OOV Precision	0.569	0.757

Table 1. Results at SIGHAN Bakeoff 2006

The performances of our system on the two corpuses can rank in the half-top group among the participated systems.

We notice that the accuracies on known word segmentation are relatively better than on OOV words segmentation. This appears somewhat unexpected. In the close experiments we had done on the PKU and MSR corpuses of SIGHAN Bakeoff 2005, the relative performance of OOV Recall was much more outstanding than of the F-measure.

We think this is due to the inappropriate parameters used in n-gram model, which over-guarantees the performance of basic word segmentation. It can be seen on the IV Recall (highest in UPUC_Close track). For only the best output sequence of the n-gram model is transferred to the HMM tagger, some potential unknown words may be miss-split in the early stage. Thus, the OOV Recall is not very good, and this also affects the overall performance.

On the other hand, the performances of OOV identification on UPUC are much better than on MSRA, while the performances of overall segmentation accuracy on UPUC are worse than on MSRA. This phenomenon also happened in our experiments on the Bakeoff 2005 corpuses of PKU and MSR. In the PKU test data, the rate of OOV words according is 0.058 while in MSR is 0.026. Thus, it can be conclude that the more unknown words occur, the more significant ability of OOV words identification appears.

In addition, the relative performance of OOV Precision are much better. This demonstrates that the OOV identification ability of our system is appreciable. In other words, the POC-NLW tagging method introduced is effective to some extent.

5 CONCLUSION AND FURTHER WORK

In this paper, a POC-NLW template is presented for word segmentation, which aims at exploring the word creation mechanisms in Chinese language by utilizing the character-level information to. A two-stage strategy was applied in our system to combine the n-gram model based word segmentation and OOV word identification implemented by a HMM tagger. Test results show that the method achieved high performance on word segmentation, especially on unknown words identification. Therefore, the method is a practical one that can be implemented as an inte-

gral component in actual Chinese NLP applications.

From the results, it can safely conclude that method introduced here does find some character-level information, and the information could effectively conduct the word segmentation and unknown words identification. For this is the first time we participate in this bakeoff, and the work has been done as a integral part of another system during the past two months, the implementation of the segmentation system we submitted is coarse. A lot of improvements, on either theoretical methods or implementation techniques, are required in our future work, including the smoothing techniques in the n-gram model and the HMM model, the refine of the features extraction method and the POC-NLW template itself, the more harmonious integration strategy and so on.

Acknowledgements

This work is partially supported by NSFC (National Natural Science Foundation of China) under Grant No.60475007, Key Project of Chinese Ministry of Education under Grant No.02029 and the Foundation of Chinese Ministry of Education for Century Spanning Talent.

References

Andi Wu, and Zixin Jiang. 2000. Statistically-enhanced new word identification in a rule-based Chinese system. *Proceedings of the 2nd Chinese Language Processing Workshop*, 46-51.

Frederick Jelinek, and Robert L. Mercer. 1980. Interpolated Estimation of Markov Source Parameters from Sparse Data. *Proceedings of Workshop on Pattern Recognition in Practice*, Amsterdam, 381-397.

Guohong Fu, and Kang-Kwong Luke. 2003. A Two-stage Statistical Word Segmentation System for Chinese. *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 156-159.

Heng Li, Yuan Dong, Xinnian Mao, Haila Wang, and Wu Liu. 2005. Chinese Word Segmentation in FTRD Beijing. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 150-153.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese Language Procession*, 8(1):29–48.

Tetsuji Nakagawa. 2004. Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. *Proceedings of the 20th International Conference on Computational Linguistics*, 466–472.

Thorsten Brants. 2000. TnT — A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000*, 224–231.