HMM Revises Low Marginal Probability by CRF for Chinese Word Segmentation*

Degen Huang, Deqin Tong, Yanyan Luo
Department of Computer Science and Engineering
Dalian University of Technology
huangdg@dlut.edu.cn, {tongdeqin, ziyanluoyu}@gmail.com

Abstract

This paper presents a Chinese word segmentation system for CIPS-SIGHAN 2010 Chinese language processing task. Firstly, based on Conditional Random Field (CRF) model, with local features and global features, the character-based tagging model is designed. Secondly, Hidden Markov Models (HMM) is used to revise the substrings with low marginal probability by CRF. Finally, confidence measure is used to regenerate the result and simple rules to deal with the strings within letters and numbers. As is well known that character-based approach has outstanding capability of discovering out-of-vocabulary (OOV) word, but external information of word lost. HMM makes use of word information to increase in-vocabulary (IV) recall. We participate in the simplified Chinese word segmentation both closed and open test on all four corpora, which belong to different domains. Our system achieves better performance.

1 Introduction

Chinese Word Segmentation (CWS) has witnessed a prominent progress in the first four SIGHAN Bakeoffs. Since Xue (2003) used character-based tagging, this method has attracted more and more attention. Some previous work (Peng et al., 2004; Tseng et al., 2005; Low et al., 2005) illustrated the effectiveness of using characters as tagging units, while literatures (Zhang et al., 2006; Zhao and Kit, 2007a; Zhang and Clark, 2007) focus on employing lexical words or subwords as tagging units. Because the word-based models can capture the word-level contextual information and IV knowledge. Besides, many strategies are proposed to balance the IV and OOV performance (Wang et al., 2008).

CRF has been widely used in sequence labeling tasks and has a good performance (Lafferty et al., 2001). Zhao and Kit (2007b; 2008) attempt to integrate global information with local information to further improve CRF-based tagging method of CWS, which provides a solid foundation for strengthening CRF learning with unsupervised learning outcomes.

In order to increase the accuracy of tagging using CRF, we adopt the strategy, which is: if the marginal probability of characters is lower than a threshold, the modified component based on HMM will be trigged; combining the confidence measure the results will be regenerated.

2 Our word segmentation system

In this section, we describe our system in more details. Three modules are included in our system: a basic character-based CRF tagger, HMM which revises the substrings with low marginal probability and confidence measure which combines them to regenerate the result. In addition, we also use some rules to deal with the strings within letters and numbers.

2.1 Character-based CRF tagger

Tag Set A 6-tag set is adopted in our system. It includes six tags: B, B2, B3, M, E and S. Here, Tag B and E stand for the first and the last position in a multi-character word, respectively. S stands for a single-character word. B2 and B3 stand for the second and the third position in a

* The work described in this paper is supported by Microsoft Research Asia Funded Project.

multi-character word. M stands for the fourth or more rear position in a multi-character word with more than four characters. The 6-tag set is proved to work more effectively than other tag sets in improving the segmentation performance of CRFs by Zhao et al. (2006).

Feature templates In our system, six n-gram templates, namely, C_{-1}, C_0, C_1, $C_{-1}C_0$, C_0C_1, $C_{-1}C_1$ are selected as features, where C stands for a character and the subscripts -1, 0 and 1 stand for the previous, current and next character, respectively. Furthermore, another one is character type feature template $T_{-1}T_0T_1$. We use four classes of character sets which are predefined as: class *N* represents numbers, class *L* represents non-Chinese letters, class *P* represents punctuation labels and class *C* represents Chinese characters.

Except for the character feature, we also employ global word feature templates. The basic idea of using global word information for CWS is to inform the supervised learner how likely it is that the subsequence can be a word candidate. The accessor variety (AV) (Feng et al., 2005) is opted as global word feature, which is integrated into CRF successfully in literatures (Zhao and Kit, 2007b; Zhao and Kit, 2008). The *AV* value of a substring s is defined as:

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

Where the left and right *AV* values $L_{av}(s)$ and $R_{av}(s)$ are defined, respectively, as the number of its distinct predecessors and the number of its distinct successors.

Multiple feature templates are used to represent word candidates of various lengths identified by the *AV* criterion. Meanwhile, in order to alleviate the sparse data problem, we follow the feature function definition for a word candidate s with a score $AV(s)$ in Zhao and Kit (2008), namely:

$$f_n(s) = t, \quad 2^t \le AV(s) < 2^{t+1} \qquad (2)$$

In order to improve the efficiency, all candidates longer than five characters are given up. The AV features of word candidates can't directly be utilized to direct CRF learning before being transferred to the information of characters. So we only choose the one with the greatest AV score to activate the above feature function for that character.

In the open test, we only add another feature of 'FRE', the basic idea of which is if a string matches a word in an existing dictionary, it may be a clue that the string is likely a true word. Then more word boundary information can be obtained, which may be helpful for CRF learning on CWS. The dictionary we used is downloaded from the Internet[①] and consists of 108,750 words with length of one to four characters. We get FRE features similar to the AV features.

2.2 HMM revises substrings with low marginal probability

The MP (short for marginal probability) of each character labeled with one of the six tags can be got separately through the basic CRF tagger. Here, B replaces 'B' and 'S' , and I represents other tags ('B_2', 'B_3', 'M', 'E'). So each character has corresponding new MP as defined in formula (3) and (4).

$$P_B = \frac{(P_S + P_B)}{\sum P_t} \qquad (3)$$

$$P_I = \frac{(P_{B_2} + P_{B_3} + P_M + P_E)}{\sum P_t} \qquad (4)$$

Where $t \in \{S, B, B_2, B_3, M, E\}$ and P_t can be calculated by using forward-backward algorithm and more details are in Lafferty et al. (2001).

A low confident word refers to a word with word boundary ambiguity which can be reflected by the MP of the first character of a word. That is, it's a low confident word if the MP of the first character of the word is lower than a threshold β (it's an empirical value and can be obtained by experiments). After getting the new MP, all these low confident candidate words are recombined with their direct predecessors until the occurrence of a word that the MP of its first character is above the threshold β, and then a new substring is generated for post processing.

Then, we use class-based HMM to re-segment the substrings mentioned above. Given a word

[①] http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information%20Processing/Source_Code/Chapter_8/Lexicon_full.zip

w_i, a word class c_i is the word itself. Let W be the word sequence, let C be its class sequence, and let $W^{\#}$ be the segmentation result with the maximum likelihood. Then, a class-based HMM model (Liu, 2004) can be got.

$$
\begin{aligned}
W^{\#} &= \arg\max_{W} P(W) \\
&= \arg\max_{W} P(W \mid C)P(C) \\
&= \arg\max_{w_1 w_2 \ldots w_m} \prod_{i=1}^{m} p'(w_i \mid c_i)P(c_i \mid c_{i-1}) \\
&= \arg\max_{w_1 w_2 \ldots w_m} \prod_{i=1}^{m} P(c_i \mid c_{i-1}) \qquad (5)
\end{aligned}
$$

Where $P(c_i \mid c_{i-1})$ indicates the transitive probability from one class to another and it can be obtained from training corpora.

The word boundary of results from HMM is also represented by tag 'B' and 'I' which meaning are the same as mentioned in above.

2.3 Confidence measure and post processing for final result

There are two segmentation results for substrings with low MP candidates after reprocessing using HMM. Analyzing experiments data, we find wrong tags labeled by CRF are mainly: OOV words in test data, IV words and incorrect words recognized by CRF. Rectifying the tags with lower MP simply may produce an even worse performance in some case. For example, some OOV words are recognized correctly by CRF but with low MP. So, we can't accept the revised results completely. A confidence measure approach is used to resolve this problem. Its calculation is defined as:

$$P_C = P_{C_o} + \lambda(1 - P_{C_o}) \qquad (6)$$

P_{C_o} is the MP of the character as 'I', λ is the premium coefficient. Based on the new value, a threshold t was used, if the value was lower than t, the original tag 'I' will be rejected and changed into the tag 'B' which is labeled by HMM.

At last, we use a simple rule to post-process the result directed at the strings that containing letters, numbers and punctuations. If the punctuation (not all punctuations) is half-width and the string before or after are composed of letters and numbers, combine all into a string as a whole. For an example, '.', '/', ':', '%' and '\' are usually recognized as split tokens. So, it needs handling additionally.

3 Experiments results and analysis

We evaluate our system on the corpora given by CIPS-SIGHAN 2010. There are four test corpora which belong to different domains. The details are showed in table 1.

Domain	Testing Data	OOV rate
A	149K	0.069
B	165K	0.152
C	151K	0.110
D	157K	0.087

Table 1. Test corpora details

A, B, C and D represent literature, computer science, medical science and finance, respectively.

3.1 Closed test

The rule for the closed test in Bakeoff is that no additional information beyond training corpora is allowed. Following the rule, the closed test is designed to compare our system with other CWS systems. Five metrics of SIGHAN Bakeoff are used to evaluate the segmentation results: F-score (F), recall (R), precision (P), the recall on IV words (R_{IV}) and the recall on OOV words (Roov). The closed test results are presented in table 2.

Domain	R	P	F	R_{oov}	R_{IV}[②]
A	0.932	0.936	0.934	0.662	0.952
	0.940	0.942	0.941	0.649	0.961
B	0.950	0.948	0.949	0.831	0.971
	0.953	0.950	0.951	0.827	0.975
C	0.934	0.932	0.933	0.751	0.957
	0.942	0.936	0.939	0.750	0.965
D	0.955	0.957	0.956	0.837	0.966
	0.959	0.960	0.959	0.827	0.972

Table 2. Evaluation closed results on all data sets

[②] In order to analyze our results, we got value of R_{IV} from the organizers because it can't be obtained from the scoring system on http://nlp.ict.ac.cn/demo/CIPS-SIGHAN2010/#.

In each domain, the first line shows the results of our basic CRF segmenter and the second one shows the final results dealt with HMM through confidence measure, which make it clear that using the confidence measure can improve the overall F-score by increasing value of R and P.

Domain	ID	R	P	F	R_{oov}	R_{IV}
A	5	0.945	0.946	0.946	0.816	0.954
	our	0.940	0.942	0.941	0.649	0.961
	12	0.937	0.937	0.937	0.652	0.958
B	our	0.953	0.950	0.951	0.827	0.975
	11	0.948	0.945	0.947	0.853	0.965
	12	0.941	0.940	0.940	0.757	0.974
C	our	0.942	0.936	0.939	0.750	0.965
	18	0.937	0.934	0.936	0.761	0.959
	5	0.940	0.928	0.934	0.761	0.962
D	our	0.959	0.960	0.959	0.827	0.972
	12	0.957	0.956	0.957	0.813	0.971
	9	0.956	0.955	0.956	0.857	0.965

Table 3. Comparison our closed results with the top three in all test sets

Next, we compare it with other top three systems. From the table 3 we can see that our system achieves better performance on closed test. In contrast, the values of R_{IV} of our method are superior to others', which contributes to the model we use. Whether the features of AV for character-based CRF tagger or HMM revising, they all make good use of word information of training corpora.

3.2 Open test

In the open test, the only additional source we use is the dictionary mentioned above. We get one first and two third best. Our result is showed in table 4. Compared with closed test, the value of R_{IV} is increased in all test corpora. But we only get the higher value of F in domain of literature. The reasons will be analyzed as follows:

In the open test, the OOV words are split into pieces because our model may be more dependent on the dictionary information. Consequently, we get higher value of R but lower P. The training corpora are the same as closed test, but it is different that FRE features are added. The additional features enhance the original information of IV words, so the value of R_{IV} is improved to some extent. However, they have side effects for OOV segmentation. We will continue to solve this problem in the future work.

Domain	R	P	F	R_{oov}	R_{IV}
A	0.956	0.947	0.952	0.636	0.980
	0.958	0.953	0.955	0.655	0.981
B	0.943	0.921	0.932	0.716	0.985
	0.948	0.929	0.939	0.735	0.986
C	0.947	0.915	0.931	0.659	0.983
	0.951	0.92	0.935	0.67	0.986
D	0.962	0.948	0.955	0.760	0.981
	0.964	0.95	0.957	0.763	0.983

Table 4. Evaluation open results on all test sets

4 Conclusions and future work

In this paper, a detailed description on a Chinese segmentation system is presented. Based on intermediate results from a CRF tagger, which employs local features and global features, we use class-based HMM to revise the substrings with low marginal probabilities. Then, a confidence measure is introduced to combine the two results. Finally, we post process the strings within letters, numbers and punctuations using simple rules. The results above show that our system achieves the state-of-the-art performance.

The MP plays the important role in our method and HMM revises some errors identified by CRF. Besides, the word features are proved to be informative cues in obtaining high quality MP. Therefore, our future work will focus on how to make CRF generate more reliable MP of characters, including exploring other word information or more unsupervised segmentation information.

References

Feng Haodi, Kang Chen, Chuyu Kit, Xiaotie Deng. 2005. *Unsupervised segmentation of Chinese corpus using accessor variety*, In: Natural Language Processing IJCNLP, pages 694-703, Sanya, China.

Lafferty John, Andrew McCallum and Fernando Pereira. 2001. *Conditional Random Fields: probabilistic models for segmenting and labeling sequence data*, In: Proceedings of ICML-18, pages 282-289, Williams College, USA.

Liu Qun, Huaping Zhang, Hongkui Yu and Xueqi Chen. 2004. *Chinese lexical analysis using cascaded Hidden Markov Model*, Journal of computer research and development 41(8): 1421-1429.

Low Kiat Jin, Hwee Tou Ng and Wenyuan Guo. 2005. *A Maximum Entropy Approach to Chinese Word Segmentation.* In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pages 161-164, Jeju Island, Korea.

Peng Fuchun, Fangfang Feng and Andrew McCallum. 2004. *Chinese segmentation and new word detection using Conditional Random Fields*, In: COLING 2004, pages 562-568, Geneva, Switzerland.

Tseng Huihsin, Pichuan Chang et al. 2005. *A Conditional Random Field Word Segmenter for SIGHAN Bakeoff 2005.* In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pages 168-171, Jeju Island, Korea.

Wang Zhenxing, Changning Huang and Jingbo Zhu. 2008. *Which perform better on in-vocabulary word segmentation: based on word or character?* In: Processing of the Sixth SIGHAN Workshop on Chinese Language Processing, pages 61-68, Hyderabad, India.

Xue Nianwen. 2003. *Chinese word segmentation as character tagging*, Computational Linguistics and Chinese Language Processing 8(1): 29-48.

Zhang Yue and Stephen Clark. 2007. *Chinese Segmentation with a Word-Based Perceptron Algorithm*. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pages 840-847, Prague, Czech Republic.

Zhang Ruiqiang, Genichiro Kikui and Eiichiro Sumita. 2006. *Subword-based tagging by Conditional Random Fields for Chinese word segmentation*, In: Proceedings of the Human Language Technology Conference of the NAACL, pages 193-196, New York, USA.

Zhao Hai, Changning Huang, Mu Li and Baoliang Lu. 2006. *Effective tag set selection in Chinese word segmentation via Conditional Random Field modeling*, In: PACLIC-20, pages 87-94, Wuhan, China.

Zhao Hai and Chunyu Kit. 2007a. *Effective subsequence based tagging for Chinese word segmentation*, Journal of Chinese Information Processing 21(5): 8-13.

Zhao Hai and Chunyu Kit. 2007b. *Incorporating global information into supervised learning for Chinese word segmentation*, In: PACLING-2007, pages 66-74, Melbourne, Australia.

Zhao Hai and Chunyu Kit. 2008. *Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition*, In: Proceedings of the Six SIGHAN Workshop on Chinese Language Processing, pages 106-111, Hyderabad, India.