Chinese Word Segmentation based on an Approach of Maximum Entropy Modeling

Yan Song[1] **Jiaqing Guo**[1] **Dongfeng Cai**[2]
Natural Language Processing Lab
Shenyang Institute of Aeronautical Engineering
Shenyang, 110034, China
1.{mattsure,guojiaqing}@gmail.com
2.cdf@ge-soft.com

Abstract

In this paper, we described our Chinese word segmentation system for the 3rd SIGHAN Chinese Language Processing Bakeoff Word Segmentation Task. Our system deal with the Chinese character sequence by using the Maximum Entropy model, which is fully automatically generated from the training data by analyzing the character sequences from the training corpus. We analyze its performance on both closed and open tracks on Microsoft Research (MSRA) and University of Pennsylvania and University of Colorado (UPUC) corpus. It is shown that we can get the results just acceptable without using dictionary. The conclusion is also presented.

1 Introduction

In the 3rd SIGHAN Chinese Language Processing Bakeoff Word Segmentation Task, we participated in both closed and open tracks on Microsoft Research corpus (MSRA for short) and University of Pennsylvania and University of Colorado corpus (UPUC for short). The following sections described how our system works and presented the results and analysis. Finally, the conclusion is presented with discussions of the system.

2 System Overview

Using Maximum Entropy approach for Chinese Word Segmentation is not a fresh idea, some previous works (Xue and Shen, 2003; Low, Ng and Guo, 2005) have got good performance in this field. But what we consider in the process of Segmentation is another way. We treat the input text which need to be segmented as a sequence of the Chinese characters, The segment process is, in fact, to find where we should split the character sequence. The point is to get the segment probability between 2 Chinese characters, which is different from dealing with the character itself.

In this section, training and segmentation process of the system is described to show how our system works.

2.1 Pre-Process of Training

For the first step we find the Minimal Segment Unit (MSU for short) of a text fragment in the training corpus. A MSU is a character or a string which is the minimal unit in a text fragment that cannot be segmented any more. According to the corpus, all of the MSUs can be divided into 5 type classes: "C" - Chinese Character (such as "你" and "好"), "AB" - alphabetic string (such as "SIGHAN"), "EN" - digit string (such as "1234567"), "CN" - Chinese number string (such as "一百二十") and "P" - punctuation ("，", "。", "；", etc). Besides the classes above, we define a tag "NL" as a special MSU, which refers to the beginning or ending of a text fragment. So, any MSU u can be described as: $u \in C \cup AB \cup EN \cup CN \cup P \cup \{NL\}$. In order to check the capability of the pure Maximum Entropy model, in closed tracks, we didn't have any type of classes, the MSU here is every character of the text fragment, $u \in C' \cup \{NL\}$. For instance, "我们参加了SIGHAN2006分词大赛。" is segmented into these MSUs: "我/们/参/加/了/S/I/G/H/A/N/2/0/0/6/分/词/大/赛/。".

Once we get all the MSUs of a text fragment, we can get the value of the Nexus Coefficient (NC for short) of any 2 adjacent MSUs according to the training corpus. The set of NC value can be

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 201–204,
Sydney, July 2006.

described as: $NC \in \{0, 1\}$, where 0 means those 2 MSUs are segmented and 1 means they are not segmented (Roughly, we appoint $r = 0$ if either one of the 2 adjacent MSUs is NL). For example, the NC value of these 2 MSUs "你" and "好" in the text fragment "你好" is 0 since these 2 characters is segmented according to the training corpus.

2.2 Training

Since the segmentation is to obtain NC value of any 2 adjacent MSUs (here we call the interspace of the 2 adjacent MSUs a check point, illustrated below),

$$\ldots \mathrm{U}_{-3}\,\mathrm{U}_{-2}\;\mathrm{U}_{-1}\;\mathrm{U}_{+1}\,\mathrm{U}_{+2}\,\mathrm{U}_{+3}\ldots$$

Check Point of U_{-1} and U_{+1}

we built a tool to extract the feature as follows:

(α) $U_{-3}, U_{-2}, U_{-1}, U_{+1}, U_{+2}, U_{+3}$
(β) $U_{-1}U_{+1}$
(γ) $r_{-2}r_{-1}$
(δ) $U_{-3}r_{-2}, U_{-2}r_{-1}$
(ϵ) $r_{-2}U_{-2}, r_{-1}U_{-1}$

In these features above, U_{+n} (U_{-n}) refers to the following (previous) *n* MSU of the check point with the information of relative position (Intuitively, We consider the same MSU has different effect on the NC value of the check point when its relative position is different to check point). And $U_{-1}U_{+1}$ is the 2 adjacent MSUs of the check point. $r_{-2}r_{-1}$ is the NC value of the previous 2 check points. Similarly, the (δ) and (ϵ) features represent the MSUs with their adjacent r. For instance, in the sentence 我是一个中国人, we can extract these features for the check point between the MSU 我 and 是:

(α) NL_{-3},NL_{-2},我$_{-1}$,是$_{+1}$,一$_{+2}$,个$_{+3}$,
(β) 我$_{-1}$是$_{+1}$
(γ) *00* (because 我 is the boundary of the sentence)
(δ) NL_{-3}*0*,NL_{-2}*0*
(ϵ) *0*NL_{-2},*0*我$_{-1}$

and also these features for the check point between the MSU 个 and 中:

(α) 是$_{-3}$,一$_{-2}$,个$_{-1}$,中$_{+1}$,国$_{+2}$,人$_{+3}$
(β) 个$_{-1}$中$_{+1}$

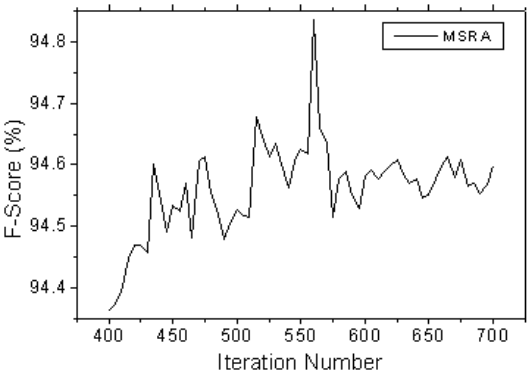

Figure 1: MSRA training curve

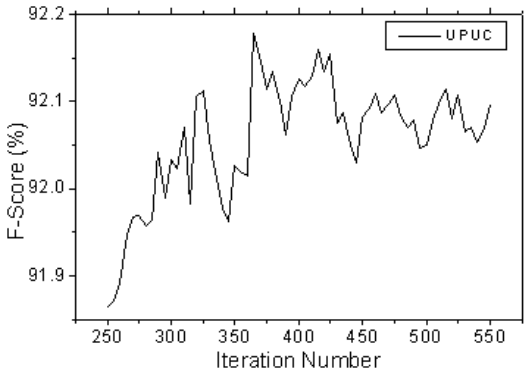

Figure 2: UPUC training curve

(γ) *01* (for UPUC corpus, here the value is 00 since 一个 is segmented into 2 characters, but in MSRA corpus, 一个 is treated as a word)
(δ) 是$_{-3}$*0*,一$_{-2}$*1*
(ϵ) *0*一$_{-2}$,*1*个$_{-1}$

After the extraction of the features, we use the ZhangLe's Maximum Entropy Toolkit[1] to train the model with a feature cutoff of 1. In order to get the best number of iteration, 9/10 of the training data is used to train the model, and the other 1/10 portion of the training data is used to evaluate the model. Figure 1 and 2 show the results of the evaluation on MSRA and UPUC corpus.

From the figures we can see the best iteration number range from 555 to 575 for MSRA corpus, and 360 to 375 for UPUC corpus. So we decide the iteration for 560 rounds for MSRA tracks and 365 rounds for UPUC tracks, respectively.

2.3 Segmentation

As we mentioned in the beginning of this section, the segmentation is the process to obtain the value

[1]Download from http://maxent.sourceforge.net

of every NC in a text fragment. This process is similar to the training process. Firstly, We scan the text fragment from start to end to get all of the MSUs. Then we can extract all of the features from the text fragment and decide which check point we should tag as $r = 0$ by this equation:

$$p(r|c) = \frac{1}{Z} \prod_{j=1}^{K} \alpha_j^{f_j(r|c)} \quad (1)$$

where K is the number of features, Z is the normalization constant used to ensure that a probability distribution results, and c represents the context of the check point. α_j is the weight for feature f_j, here $\{\alpha_1\alpha_2 \ldots \alpha_K\}$ is generated by the training data. We then compute $P(r = 0|c)$ and $P(r = 1|c)$ by the equation (1).

After one check point is treated with value of r, the system shifts backward to the next check point until all of the check point in the whole text fragment are treated. And by calculating:

$$P = \prod_{i=1}^{n-1} p(r_i|c_i) = \prod_{i=1}^{n-1} \frac{1}{Z} \prod_{j=1}^{K} \alpha_j^{f_k(r_i|c_i)} \quad (2)$$

to get an r sequence which can maximize P. From this process we can see that the sequence is, in fact, a second-order Markov Model. Thus it is easily to think about more tags prior to the check point (as an n^{th}-order Markov Model) to get more accuracy, but in this paper we only use the previous 2 tags from the check point.

2.4 Identification of New words

We perform the new word(s) identification as a post-process by check the word formation power (WFP) of characters. The WFP of a character is defined as: $WFP(c) = N_{wc}/N_c$, where N_{wc} is the number of times that the character c appears in a word of at least 2 characters in the training corpus, N_c is the number of times the character c occurs in the training corpus. After a text fragment is segmented by our system, we extract all consecutive single characters. If at least 2 consecutive characters have the WFP larger than our threshold of 0.88, we polymerize them together as a word. For example, “州务卿” is a new word which is segmented as “州/务/卿” by our system, WFP of these 3 characters is 0.9517,0.9818 and 1.0 respectively, then they are polymerized as one word.

Besides the WFP, during the experiments, we find that the Maximum Entropy model can polymerize some MSUs as a new word (We call it polymerization characteristic of the model), such as 闻风而动 in the training corpus, we can extract 闻风而 as the previous context feature of the check point after 而, in another string 嘎然而止, we can extract the backward context 止 of the check point after 而 with $r = 1$. Then in the test, a new word 闻风而止 is recognized by the model since 闻风而 and 止 are polymerized if 而止 appears together a large number of times in the training corpus.

3 Performance analysis

Here Table 1 illustrates the results of all 4 tracks we participate. The first column is the track name, and the 2nd column presents the Recall (R), the 3rd column the Precision (P), the 4th column is F-measure (F). The R_{oov} refers to the recall of the out-of-vocabulary words and the R_{iv} refers to the recall of the words in training corpus.

Track	R	P	F	R_{oov}	R_{iv}
MSRA Closed	0.923	0.929	0.926	0.554	0.936
MSRA Open	0.938	0.946	0.942	0.706	0.946
UPUC Closed	0.902	0.887	0.895	0.568	0.934
UPUC Open	0.926	0.906	0.917	0.660	0.954

Table 1: Results of our system in 4 tracks.

3.1 Closed tracks

For all of the closed tracks, we perform the segmentation as we mentioned in the section above, without any class defined. Every MSU we extract from the training data is a character, which may be a Chinese character, an English letter or a single digit. We extract the features based on this kind of MSUs to generate the models. The results show these models are not precise.

For the UPUC closed track, the official released training data is rather small. Then the capability of the model is limited, this is the most reasonable negative effect on our F-measure 0.895.

3.2 Open tracks

The primary change between open tracks and closed tracks is that we have classified 5 classes (“C”,“AB”,“EN”,“CN” and “P”) to MSUs in order to improve the accuracy of the model. The classification really works and affects the performance of the system in a great deal. As this text fragment 1998年 can be recognized as (EN)(C), which can also presents 1644年, thus 1644年 can

be easily recognized though there is no 1664年 in the training data.

The training corpus we used in UPUC open track is the same as in UPUC closed track. With those 5 classes, it is easily seen that the F-measure increased by 2.2% in the open tracks.

For the MSRA open track, we adjust the class "P" by removing the punctuation "、" from the class, because in the MSRA corpus, "、" can be a part of a organization name, such as "、" in "中俄友好、和平与发展委员会". Besides, we add the Microsoft Research training data of SIGHAN bakeoff 2005 as extended training corpus. The larger training data cooperate with the classification method, the F-measure of the open track increased to 0.942 as comparison with 0.926 of closed track.

3.3 Discussion of the tracks

Through the tracks, we tested the performance by using the pure Maximum Entropy model in closed tracks and run with the improved model with classified MSUs in open tracks. It is shown that the pure model without any additional methods can hardly make us satisfied, for the open tracks, the model with classes are just acceptable in segmentation.

In both closed and open tracks, we use the same new word identification process, and with the polymerization characteristic of the model, we find the R_{oov} is better than we expected.

On the other hand, in our system, there is no dictionary used as we described in the sections above, the R_{iv} of each track shows that affects the system performance.

Another factor affects our system in the UPUC tracks is the wrongly written characters. Consider that our system is based on the sequence of characters, this kind of mistake is fatal. For example, in the sentence 他们无愧于最可爱的人的美喻, where 美誉 is written as 美喻. The model cannot recognize it since 美喻 didn't occur in the training corpus. In the step of new word identification, the WFPs of the 2 characters 美，喻 are 0.8917 and 0.8310, thus they are wrongly segmented into 2 single characters while they are treated as a word in the gold standard corpus. Therefore, we believe the results can increase if there are no such mistakes in the test data.

4 Conclusion

We propose an approach to Chinese word segmentation by using Maximum Entropy model, which focuses on the nexus relationship of any 2 adjacent MSUs in a text fragment. We tested our system with pure Maximum Entropy models and models with simplex classification method. Compare with the pure models, the models with classified MSUs show us better performances. However, the Maximum Entropy models of our system still need improvement if we want to achieve higher performance. In future works, we will consider using more training data and add some hybrid methods with pre- and post-processes to improve the system.

Acknowledgements

We would like to thank all the colleagues of our Lab. Without their encouragement and help, this work cannot be accomplished in time.

This is our first time to participate such an international bakeoff. There are a lot of things we haven't experienced ever before, but with the enthusiastic help from the organizers, we can come through the task. Especially, We wish to thank Gina-Anne Levow for her patience and immediate reply for any of our questions, and we also thank Olivia Kwong for the advice of paper submission.

References

Nianwen Xue and Libin Shen. 2003. *Chinese Word Segmentation as LMR tagging*. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, p176-179.

Maosong Sun, Ming Xiao, B K Tsou. 2004. *Chinese Word Segmentation without Using Dictionary Based on Unsupervised Learning Strategy*. Chinese Journal of Computers, Vol.27, #6, p736-742.

Jin Kiat Low, 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, p161-164.