

A Maximum Entropy Approach to Chinese Word Segmentation

Jin Kiat Low¹ and Hwee Tou Ng^{1,2} and Wenyuan Guo²

1. Department of Computer Science, National University of Singapore,
3 Science Drive 2, Singapore 117543

2. Singapore-MIT Alliance, E4-04-10, 4 Engineering Drive 3, Singapore 117576
{lowjinki, nght, guowy}@comp.nus.edu.sg

Abstract

We participated in the Second International Chinese Word Segmentation Bakeoff. Specifically, we evaluated our Chinese word segmenter in the open track, on all four corpora, namely Academia Sinica (AS), City University of Hong Kong (CITYU), Microsoft Research (MSR), and Peking University (PKU). Based on a maximum entropy approach, our word segmenter achieved the highest F measure for AS, CITYU, and PKU, and the second highest for MSR. We found that the use of an external dictionary and additional training corpora of different segmentation standards helped to further improve segmentation accuracy.

1 Chinese Word Segmenter

The Chinese word segmenter we built is similar to the maximum entropy word segmenter we employed in our previous work (Ng and Low, 2004). Our word segmenter uses a maximum entropy framework (Ratnaparkhi, 1998; Xue and Shen, 2003) and is trained on manually segmented sentences. It classifies each Chinese character given the features derived from its surrounding context. Each Chinese character can be assigned one of four possible boundary tags: *s* for a character that occurs as a single-character word, *b* for a character that begins a multi-character (i.e., two or more characters) word, *e* for a character that ends a multi-character word, and *m* for a character that is

neither the first nor last in a multi-character word. Our implementation used the opennlp maximum entropy package v2.1.0 from sourceforge.¹

1.1 Basic Features

The basic features of our word segmenter are similar to our previous work (Ng and Low, 2004):

- (a) $C_n (n = -2, -1, 0, 1, 2)$
- (b) $C_n C_{n+1} (n = -2, -1, 0, 1)$
- (c) $C_{-1} C_1$
- (d) $Pu(C_0)$
- (e) $T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

In the above feature templates, C refers to a Chinese character. Templates (a) – (c) refer to a context of five characters (the current character and two characters to its left and right). C_0 denotes the current character, $C_n (C_{-n})$ denotes the character n positions to the right (left) of the current character. For example, given the character sequence “新华社北京”, when considering the character C_0 “社”, C_{-2} denotes “新”, $C_1 C_2$ denotes “北京”, etc. The punctuation feature, $Pu(C_0)$, checks whether C_0 is a punctuation symbol (such as “?”, “-”, “,”). For the type feature (e), four type classes are defined: numbers represent class 1, dates (“日”, “月”, “年”, the Chinese characters for “day”, “month”, “year”, respectively) represent class 2, English letters represent class 3, and other characters represent class 4. For example, when considering the character “年” in the character sequence “九零年代R”, the feature $T(C_{-2}) \dots T(C_2) = 11243$

¹<http://maxent.sourceforge.net/>

will be set to 1 (“九” is the Chinese character for “9” and “零” is the Chinese character for “0”).

Besides these basic features, we also made use of character normalization. We note that characters like punctuation symbols and Arabic digits have different character codes in the ASCII, GB, and BIG5 encoding standard, although they mean the same thing. For example, comma “,” is represented as the hexadecimal value 0x2c in ASCII, but as the hexadecimal value 0xa3ac in GB. In our segmenter, these different character codes are normalized and replaced by the corresponding character code in ASCII. Also, all Arabic digits are replaced by the ASCII digit “0” to denote any digit. Incorporating character normalization enables our segmenter to be more robust against the use of different encodings to represent the same character.

For all the experiments that we conducted, training was done with a feature cutoff of 2 and 100 iterations, except for the AS corpus which had a feature cutoff of 3.

A major difficulty faced by a Chinese word segmenter is the presence of out-of-vocabulary (OOV) words. Segmenting a text with many OOV words tends to result in lower accuracy. We address the problem of OOV words in two ways: using an external dictionary containing a list of predefined words, and using additional training corpora which are not segmented according to the same segmentation standard.

1.2 External Dictionary

If a sequence of characters in a sentence matches a word in an existing dictionary, it may be a clue that the sequence of characters should be segmented as one word. We used an online dictionary from Peking University downloadable from the Internet², consisting of about 108,000 words of length one to four characters. If there is some sequence of neighboring characters around C_0 in the sentence that matches a word in this dictionary, then we greedily choose the longest such matching word W in the dictionary. Let t_0 be the boundary tag of C_0 in W , L the number of characters in W , and $C_1(C_{-1})$ be the character

immediately following (preceding) C_0 in the sentence. We then add the following features derived from the dictionary:

- (f) Lt_0
- (g) $C_nt_0(n = -1, 0, 1)$

For example, consider the sentence “新华社北京...”. When processing the current character C_0 “华”, we will attempt to match the following candidate sequences “华”, “新华”, “华社”, “新华社”, “华社北”, “新华社北”, and “华社北京” against existing words in our dictionary. Suppose both “华社” and “新华社” are found in the dictionary. Then the longest matching word W chosen is “新华社”, t_0 is m , L is 3, C_{-1} is “新”, and C_1 is “社”.

1.3 Additional Training Corpora

The presence of different standards in Chinese word segmentation limits the amount of training corpora available for the community, due to different organizations preparing training corpora in their own standards. Indeed, if one uniform segmentation standard were adopted, more training data would have been available, and the OOV problem could be significantly reduced.

We observed that although different segmentation standards exist, the differences are limited, and many words are still segmented in the same way across two different segmentation standards. As such, in our work, we attempt to incorporate corpora from other segmentation standards as additional training data, to help reduce the OOV problem.

Specifically, the steps taken are:

1. Perform training with maximum entropy modeling using the original training corpus D_0 annotated in a given segmentation standard.
2. Use the trained word segmenter to segment another corpus D annotated in a different segmentation standard.
3. Suppose a Chinese character C in D is assigned a boundary tag t by the word segmenter with probability p . If t is identical to the boundary tag of C in the gold-standard

²http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information%20Processing/Source_Code/Chapter_8/Lexicon_full_2000.zip

annotated corpus D , and p is less than some threshold θ , then C (with its surrounding context in D) is used as additional training data.

4. Add all such characters C as additional training data to the original training corpus D_0 , and train a new word segmenter using the enlarged training data.
5. Evaluate the accuracy of the new word segmenter on the same test data annotated in the original segmentation standard of D_0 .

For the current bakeoff, when training a word segmenter on a particular training corpus, the additional training corpora are all the three corpora in the other segmentation standards. For example, when training a word segmenter for the AS corpus, the additional training corpora are CITYU, MSR, and PKU. The necessary character encoding conversion between GB and BIG5 is performed, and the probability threshold θ is set to 0.8. We found from our experiments that setting θ to a higher value did not further improve segmentation accuracy, but would instead increase the training set size and incur longer training time.

2 Testing

During testing, the probability of a boundary tag sequence assignment $t_1 \dots t_n$ given a character sequence $C_1 \dots C_n$ is determined by using the maximum entropy classifier to compute the probability that a boundary tag t_i is assigned to each individual character C_i . If we were to just assign each character the boundary tag with the highest probability, it is possible that the classifier produces a sequence of invalid tags (e.g., m followed by s). To eliminate such possibilities, we implemented a dynamic programming algorithm which considers only valid boundary tag sequences given an input character sequence. At each character position i , the algorithm considers each last word candidate ending at position i and consisting of K characters in length ($K = 1, \dots, 20$ in our experiments). To determine the boundary tag assignment to the last word W with K characters, the first character of W is assigned boundary tag b , the last character of W is assigned tag e , and the intervening characters

Corpus	R	P	F	R_{OOV}	R_{IV}
AS	0.962	0.950	0.956	0.684	0.975
CITYU	0.967	0.956	0.962	0.806	0.980
MSR	0.969	0.968	0.968	0.736	0.975
PKU	0.968	0.969	0.969	0.838	0.976

Table 1: Our official SIGHAN bakeoff results

are assigned tag m . (If W is a single-character word, then the single character is assigned tag s). In this way, the dynamic programming algorithm only considers valid tag sequences.

After word segmentation is done by the maximum entropy classifier, a post-processing step is applied to correct inconsistently segmented words made up of 3 or more characters. A word W is defined to be inconsistently segmented if the concatenation of 2 to 6 consecutive words elsewhere in the segmented output document matches W . In the post-processing step, the segmentation of the characters of these consecutive words is changed so that they are segmented as a single word. To illustrate, if the concatenation of 2 consecutive words “香格” and “里拉” in the segmented output document matches another word “香格里拉”, then the 2 consecutive words “香格” and “里拉” will be re-segmented as a single word “香格里拉”.

3 Evaluation Results

We evaluated our Chinese word segmenter in the open track, on all 4 corpora, namely Academia Sinica (AS), City University of Hong Kong (CITYU), Microsoft Research (MSR), and Peking University (PKU). Table 1 shows our official SIGHAN bakeoff results. The columns R, P, and F show the recall, precision, and F measure, respectively. The columns R_{OOV} and R_{IV} show the recall on out-of-vocabulary words and in-vocabulary words, respectively. Our Chinese word segmenter which participated in the bakeoff was trained with the basic features (Section 1.1), and made use of the external dictionary (Section 1.2) and additional training corpora (Section 1.3). Our word segmenter achieved the highest F measure for AS, CITYU, and PKU, and the second highest for MSR.

After the release of the official bakeoff results,

Corpus	V1	V2	V3	V4
AS	0.953	0.955	0.956	0.956
CITYU	0.950	0.960	0.961	0.962
MSR	0.960	0.968	0.963	0.968
PKU	0.948	0.965	0.956	0.969

Table 2: Word segmentation accuracy (F measure) of different versions of our word segmenter

we ran a series of experiments to determine the contribution of each component of our word segmenter, using the official scorer and test sets with gold-standard segmentations. Version V1 used only the basic features (Section 1.1); Version V2 used the basic features and additional features derived from our external dictionary (Section 1.2); Version V3 used the basic features but with additional training corpora (Section 1.3); and Version V4 is our official submitted version combining basic features, external dictionary, and additional training corpora. Table 2 shows the word segmentation accuracy (F measure) of the different versions of our word segmenter, when tested on the official test sets of the four corpora. The results indicate that the use of external dictionary increases segmentation accuracy. Similarly, the use of additional training corpora of different segmentation standards also increases segmentation accuracy.

4 Conclusion

Using a maximum entropy approach, our Chinese word segmenter achieves state-of-the-art accuracy, when evaluated on all four corpora in the open track of the Second International Chinese Word Segmentation Bakeoff. The use of an external dictionary and additional training corpora of different segmentation standards helps to further improve segmentation accuracy.

Acknowledgements

This research is partially supported by a research grant R252-000-125-112 from National University of Singapore Academic Research Fund, as well as the Singapore-MIT Alliance.

References

- Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 277–284.
- Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.
- Nianwen Xue and Libin Shen. 2003. Chinese word segmentation as LMR tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 176–179.