Voting between Dictionary-based and Subword Tagging Models for Chinese Word Segmentation

Dong Song and **Anoop Sarkar**
School of Computing Science, Simon Fraser University
Burnaby, BC, Canada V5A1S6
{dsong,anoop}@cs.sfu.ca

Abstract

This paper describes a Chinese word segmentation system that is based on majority voting among three models: a forward maximum matching model, a conditional random field (CRF) model using maximum subword-based tagging, and a CRF model using minimum subword-based tagging. In addition, it contains a post-processing component to deal with inconsistencies. Testing on the closed track of CityU, MSRA and UPUC corpora in the third SIGHAN Chinese Word Segmentation Bakeoff, the system achieves a F-score of 0.961, 0.953 and 0.919, respectively.

1 Introduction

Tokenizing input text into words is the first step of any text analysis task. In Chinese, a sentence is written as a string of characters, to which we shall refer by their traditional name of *hanzi*, without separations between words. As a result, before any text analysis on Chinese, word segmentation task has to be completed so that each word is "isolated" by the word-boundary information.

Participating in the third SIGHAN Chinese Word Segmentation Bakeoff in 2006, our system is tested on the closed track of CityU, MSRA and UPUC corpora. The sections below provide a detailed description of the system and our experimental results.

2 System Description

In our segmentation system, a hybrid strategy is applied (Figure 1): First, forward maximum matching (Chen and Liu, 1992), which is a dictionary-based method, is used to generate a segmentation result. Also, the CRF model using maximum subword-based tagging (Zhang et al., 2006) and the CRF model using minimum subword-based tagging, both of which are statistical methods, are used individually to solve the problem. In the next step, the solutions from these three methods are combined via the *hanzi*-level majority voting algorithm. Then, a post-processing procedure is applied in order to to get the final output. This procedure merges adjoining words to match the dictionary entries and then splits words which are inconsistent with entries in the training corpus.

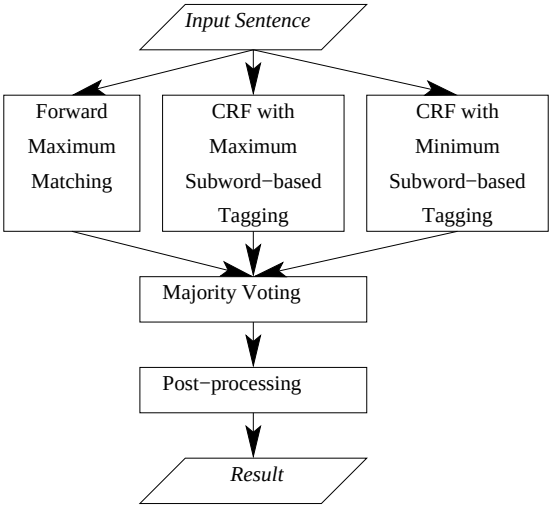

Figure 1: Outline of the segmentation process

2.1 Forward Maximum Matching

The maximum matching algorithm is a greedy segmentation approach. It proceeds through the sentence, mapping the longest word at each point with an entry in the dictionary. In our system, the well-known forward maximum matching algorithm (Chen and Liu, 1992) is implemented.

The maximum matching approach is simple and efficient, and it results in high in-vocabulary accuracy; However, the small size of the dictionary, which is obtained only from the training data, is a major bottleneck for this approach to be applied by itself.

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 126–129,
Sydney, July 2006.

2.2 CRF Model with Maximum Subword-based Tagging

Conditional random fields (CRF), a statistical sequence modeling approach (Lafferty et al., 2001), has been widely applied in various sequence learning tasks including Chinese word segmentation. In this approach, most existing methods use the character-based IOB tagging. For example, "都(all) 至关重要(extremely important)" is labeled as "都(all)/O 至(until)/B 关(close)/I 重(heavy)/I 要(demand)/I".

Recently (Zhang et al., 2006) proposed a maximum subword-based IOB tagger for Chinese word segmentation, and our system applies their approach which obtains a very high accuracy on the shared task data from previous SIGHAN competitions. In this method, all single-*hanzi* words and the top frequently occurring multi-*hanzi* words are extracted from the training corpus to form the lexicon subset. Then, each word in the training corpus is segmented for IOB tagging, with the forward maximum matching algorithm, using the formed lexicon subset as the dictionary. In the above example, the tagging labels become "都(all)/O 至(until)/B 关(close)/I 重 要(important)/I", assuming that "重要(important)" is the longest subword in this word, and it is one of the top frequently occurring words in the training corpus.

After tagging the training corpus, we use the package CRF++[1] to train the CRF model. Suppose w_0 represents the current word, w_{-1} is the first word to the left, w_{-2} is the second word to the left, w_1 is the first word to the right, and w_2 is the second word to the right, then in our experiments, the types of unigram features used include w_0, w_{-1}, w_1, w_{-2}, w_2, w_0w_{-1}, w_0w_1, $w_{-1}w_1$, $w_{-2}w_{-1}$, and w_2w_0. In addition, only combinations of previous observation and current observation are exploited as bigram features.

2.3 CRF Model with Minimum Subword-based Tagging

In our third model, we applies a similar approach as in the previous section. However, instead of finding the maximum subwords, we explore the minimum subwords. At the beginning, we build the dictionary using the whole training corpus. Then, for each word in the training data, a forward shortest matching is used to get the sequence of minimum-length subwords, and this sequence is tagged in the same IOB format as before. Suppose "a", "ac", "de" and "acde" are the only entries in the dictionary. Then, for the word "acde", the sequence of subwords is "a", "c" and "de", and the tags assigned to "acde" are "a/B c/I de/I".

After tagging the training data set, CRF++ package is executed again to train this type of model, using the identical unigram and bigram feature sets that are used in the previous model. Meanwhile, the unsegmented test data is segmented by the forward shortest matching algorithm. After this initial segmentation process, the result is fed into the trained CRF model for re-segmentation by assigning IOB tags.

2.4 Majority Voting

Having the segmentation results from the above three models in hand, in this next step, we adopt the *hanzi*-level majority voting algorithm. First, for each *hanzi* in a segmented sentence, we tag it either as "B" if it is the first *hanzi* of a word or a single-*hanzi* word, or as "I" otherwise. Then, for a given *hanzi* in the results from those three models, if at least two of the models provide the identical tag, it will be assigned that tag. For instance, suppose "a c de" is the segmentation result via forward maximum matching, and it is also the result from CRF model with maximum subword-based tagging, and "ac d e" is the result from the third model. Then, for "a", since all of them assign "B' to it, "a" is given the "B" tag; for "c", because two of segmentations tag it as "B", "c" is given the "B" tag as well. Similarly, the tag for each remaining *hanzi* is determined by this majority voting process, and we get "a c de" as the result for this example.

To test the performance of each of the three models and that of the majority voting, we divide the MSRA corpus into training set and held-out set. Throughout all the experiments we conducted, we discover that those two CRF models perform much better than the pure *hanzi*-based CRF method, and that the voting process improves the performance further.

2.5 Post-processing

While analyzing errors with the segmentation result from the held-out set, we find two inconsistency problems: First, the inconsistency between the dictionary and the result: that is, certain words that appear in the dictionary are separated into consecutive words in the test result; Second,

[1] available from http://www/chasen.org/∼taku/software

the inconsistency among words in the dictionary; For instance, both "科学研究"(scientific research) and "科学(science) 研究(research)" appear in the training corpus.

To deal with the first phenomena, for the segmented result, we try to merge adjoining words to match the dictionary entries. Suppose "a b c de" are the original voting result, and "ab", "abc" and "cd" form the dictionary. Then, we merge "a", "b" and "c" together to get the longest match with the dictionary. Therefore, the output is "abc de".

For the second problem, we introduce the *split* procedure. In our system, we only consider two consecutive words. First, all bigrams are extracted from the training corpus, and their frequencies are counted. After that, for example, if "a b" appears more often than "ab", then whenever in the test result we encounter "ab", we split it into "a b".

The post-processing steps detailed above attempt to maximize the value of known words in the training data as well as attempting to deal with the word segmentation inconsistencies in the training data.

3 Experiments and Analysis

The third International Chinese Language Processing Bakeoff includes four different corpora, Academia Sinica (CKIP), City University of Hong Kong (CityU), Microsoft Research (MSRA), and University of Pennsylvania and University of Colorado, Boulder (UPUC), for the word segmentation task.

In this bakeoff, we test our system in CityU, MSRA and UPUC corpora, and follow the closed track. That is, we only use training material from the training data for the particular corpus we are testing on. No other material or any type of external knowledge is used, including part-of-speech information, externally generated word-frequency counts, Arabic and Chinese numbers, feature characters for place names and common Chinese surnames.

3.1 Results on SIGHAN Bakeoff 2006

To observe the result of majority voting and the contribution of the post-processing step, the experiment is ran for each corpus by first producing the outcome of majority voting and then producing the output from the post-processing. In each experiment, the precision (P), recall (R), F-measure (F), Out-of-Vocabulary rate (OOV), OOV recall rate (R_{OOV}), and In-Vocabulary rate (R_{IV}) are recorded. Table 1,2,3 show the scores for the CityU corpus, for the MSRA corpus, and for the UPUC corpus, respectively.

	Majority Voting	Post-processing
P	0.956	0.958
R	0.962	0.963
F	0.959	0.961
OOV	0.04	0.04
R_{OOV}	0.689	0.689
R_{IV}	0.974	0.974

Table 1: Scores for CityU corpus

	Majority Voting	Post-processing
P	0.952	0.954
R	0.952	0.952
F	0.952	0.953
OOV	0.034	0.034
R_{OOV}	0.604	0.604
R_{IV}	0.964	0.964

Table 2: Scores for MSRA corpus

	Majority Voting	Post-processing
P	0.908	0.909
R	0.927	0.929
F	0.918	0.919
OOV	0.088	0.088
R_{OOV}	0.628	0.628
R_{IV}	0.956	0.958

Table 3: Scores for UPUC corpus

From those tables, we can see that a simple majority voting algorithm produces accuracy that is higher than each individual system and reasonably high F-scores overall. In addition, the post-processing step indeed helps to improve the performance.

3.2 Error analysis

The errors that occur in our system are mainly due to the following three factors:

First, there is inconsistency between the gold segmentation and the training corpus. Although the inconsistency problem within the training corpus is intended to be tackled in the post-processing step, we cannot conclude that the segmentation

for certain words in the gold test set always follows the convention in the training data set. For example, in the MSRA training corpus, "中国政府"(Chinese government) is usually considered as a single word; while in the gold test set, it is separated as two words "中国"(Chinese) and "政府"(government). This inconsistency issue lowers the system performance. This problem, of course, affects all competing systems.

Second, we don't have specific steps to deal with words with postfixes such as "者"(person). Compared to our system, (Zhang, 2005) proposed a segmentation system that contains morphologically derived word recognition post-processing component to solve this problem. Lacking of such a step prevents us from identifying certain types of words such as "劳动者"(worker) to be a single word.

In addition, the unknown words are still troublesome because of the limited size of the training corpora. In the class of unknown words, we encounter person names, numbers, dates, organization names and words translated from languages other than Chinese. For example, in the produced CityU test result, the translated person name "米哈伊洛维奇"(Mihajlovic) is incorrectly separated as "米哈伊洛" and "维奇". Moreover, in certain cases, person names can also create ambiguity. Take the name "秋北方"(Qiu, Beifang) in UPUC test set for example, without understanding the meaning of the whole sentence, it is difficult even for human to determine whether it is a person name or it represents "秋"(autumn), "北方"(north), with the meaning of "the autumn in the north".

4 Alternative to Majority Voting

In designing the voting procedure, we also attempt to develop and use a segmentation lattice, which proceeds using a similar underlying principle as the one applied in (Xu et al., 2005).

In our approach, for an input sentence, the segmentation result using each of our three models is transformed into an individual lattice. Also, each edge in the lattice is assigned a particular weight, according to certain features such as whether or not the output word from that edge is in the dictionary. After building the three lattices, one for each model, we merge them together. Then, the shortest path, referring to the path that has the minimum weight, is extracted from the merged lattice, and therefore, the segmentation result is determined by this shortest path.

However, in the time we had to run our experiments on the test data, we were unable to optimize the edge weights to obtain high accuracy on some held-out set from the training corpora. So instead, we tried a simple method for finding edge weights by uniformly distributing the weight for each feature; Nevertheless, by testing on the shared task data from the 2005 SIGHAN bakeoff, the performance is not competitive, compared to our simple majority voting method described above. As a result, we decide to abandon this approach for this year's SIGHAN bakeoff.

5 Conclusion

Our Chinese word segmentation system is based on majority voting among the initial outputs from forward maximum matching, from a CRF model with maximum subword-based tagging, and from a CRF model with minimum subword-based tagging. In addition, we experimented with various steps in post-processing which effectively boosted the overall performance.

In future research, we shall explore more sophisticated ways of voting, including the continuing investigation on the segmentation lattice approach. Also, more powerful methods on how to accurately deal with unknown words, including person and place names, without external knowledge, will be studied as well.

References

Keh-jiann Chen, and Shing-Huan Liu. 1992. Word Identification for Mandarin Chinese Sentences. In *Fifth International Conference on Computational Linguistics*, pages 101–107.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of ICML-2001*, pages 591–598.

Jia Xu, Evgeny Matusov, Richard Zens, and Hermann Ney. 2005. Integrated Chinese Word Segmentation in Statistical Machine Translation. In *Proc. of IWSLT-2005*.

Huipeng Zhang, Ting Liu, Jinshan Ma, and Xiantao Liu. 2005. Chinese Word Segmentation with Multiple Postprocessors in HIT-IRLab. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 172–175.

Ruiqiang Zhang, Genichiro Kikui, and Eiichiro Sumita. 2006. Subword-based Tagging by Conditional Random Fields for Chinese Word Segmentation. In *Proc. of HLT-NAACL 2006*.