A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005

Huihsin Tseng
Dept. of Linguistics
University of Colorado
Boulder, CO 80302
tseng@colorado.edu

Pichuan Chang, Galen Andrew,
Daniel Jurafsky, Christopher Manning
Stanford Natural Language Processing Group
Stanford University
Stanford, CA 94309
{pichuan, pupochik, jurafsky, manning}@stanford.edu

Abstract

We present a Chinese word segmentation system submitted to the closed track of Sighan bakeoff 2005. Our segmenter was built using a conditional random field sequence model that provides a framework to use a large number of linguistic features such as character identity, morphological and character reduplication features. Because our morphological features were extracted from the training corpora automatically, our system was not biased toward any particular variety of Mandarin. Thus, our system does not overfit the variety of Mandarin most familiar to the system's designers. Our final system achieved a F-score of 0.947 (AS), 0.943 (HK), 0.950 (PK) and 0.964 (MSR).

1 Introduction

The 2005 Sighan Bakeoff included four different corpora, Academia Sinica (AS), City University of Hong Kong (HK), Peking University (PK), and Microsoft Research Asia (MSR), each of which has its own definition of a word.

In the 2003 Sighan Bakeoff (Sproat & Emerson 2003), no single model performed well on all corpora included in the task. Rather, systems tended to do well on corpora largely drawn from a set of similar Mandarin varieties to the one they were originally developed for. Across corpora, variation is seen in both the lexicons and also in the word segmentation standards. We concluded that, for future systems, generalization across such different Mandarin varieties is crucial. To this end, we proposed a new model using character identity, morphological and character reduplication features in a conditional random field modeling framework.

2 Algorithm

Our system builds on research into conditional random field (CRF), a statistical sequence modeling framework first introduced by Lafferty et al. (2001). Work by Peng et al. (2004) first used this framework for Chinese word segmentation by treating it as a binary decision task, such that each character is labeled either as the beginning of a word or the continuation of one. Gaussian priors were used to prevent overfitting and a quasi-Newton method was used for parameter optimization.

The probability assigned to a label sequence for a particular sequence of characters by a CRF is given by the equation below:

$$P_\lambda(Y \mid X) = \frac{1}{Z(X)} \exp\left(\sum_{c \in C} \sum_k \lambda_k f_k \ (Y_c, X, c) \right)$$

Y is the label sequence for the sentence, X is the sequence of unsegmented characters, Z(X) is a normalization term, f_k is a feature function, and c indexes into characters in the sequence being labeled.

A CRF allows us to utilize a large number of n-gram features and different state sequence

based features and also provides an intuitive framework for the use of morphological features.

3 Feature engineering

3.1 Features

The linguistic features used in our model fall into three categories: character identity *n*-grams, morphological and character reduplication features.

For each state, the character identity features (Ng & Low 2004, Xue & Shen 2003, Goh et al. 2003) are represented using feature functions that key off of the identity of the character in the current, proceeding and subsequent positions. Specifically, we used four types of unigram feature functions, designated as C_0 (current character), C_1 (next character), C_{-1} (previous character), C_{-2} (the character two characters back). Furthermore, four types of bi-gram features were used, and are notationally designated here as conjunctions of the previously specified unigram features, C_0C_1, $C_{-1}C_0$, $C_{-1}C1$, $C_{-2}C_{-1}$, and C_2C_0.

Given that unknown words are normally more than one character long, when representing the morphological features as feature functions, such feature functions keyed off the morphological information extracted from both the proceeding state and the current state. Our morphological features are based upon the intuition regarding unknown word features given in Gao et al. (2004). Specifically, their idea was to use productive affixes and characters that only occurred independently to predict boundaries of unknown words. To construct a table containing affixes of unknown words, rather than using threshold-filtered affix tables in a separate unknown word model as was done in Gao et al. (2004), we first extracted rare words from a corpus and then collected the first and last characters to construct the prefix and suffix tables. For the table of individual character words, we collected an individual character word table for each corpus of the characters that always occurred alone as a separate word in the given corpus. We also collected a list of bi-grams from each training corpus to distinguish known strings from unknown. Adopting all the features together in a model and using the automatically generated morphological tables prevented our system from manually overfitting the Mandarin varieties we are most familiar with.

The tables are used in the following ways:

1) $C_{-1}+C_0$ unknown word feature functions were created for each specific pair of characters in the bi-gram tables. Such feature functions are active if the characters in the respective states match the corresponding feature function's characters. These feature functions are designed to distinguish known strings from unknown.

2) C_{-1}, C_0, and C_1 individual character feature functions were created for each character in the individual character word table, and are likewise active if the respective character matches the feature function's character.

3) C_{-1} prefix feature functions are defined over characters in the prefix table, and fire if the character in the proceeding state matches the feature function's character.

4) C_0 suffix feature functions are defined over suffix table characters, and fire if the character in the current state matches the feature function's character.

Additionally, we also use reduplication feature functions that are active based on the repetition of a given character. We used two such feature functions, one that fires if the previous and the current character, C_{-1} and C_0, are identical and one that does so if the subsequent and the previous characters, C_{-1} and C_1, are identical.

Most features appeared in the first-order templates with a few of character identity features in the both zero-order and first-order templates. We also did normalization of punctuations due to the fact that Mandarin has a huge variety of punctuations.

Table 1 shows the number of data features and lambda weights in each corpus.

Table 1 The number of features in each corpus

	# of data features	*# of lambda weights*
AS	2,558,840	8,076,916
HK	2,308,067	7,481,164
PK	1,659,654	5,377,146
MSR	3,634,585	12,468,890

3.2 Experiments

3.2.1 Results on Sighan bakeoff 2003

Experiments done while developing this system showed that its performance was significantly better than that of Peng et al. (2004).

As seen in Table 2, our system's F-score was 0.863 on CTB (Chinese Treebank from Univer-

sity of Pennsylvania) versus 0.849 F on Peng et al. (2004). We do not at present have a good understanding of which aspects of our system give it superior performance.

Table 2 Comparisons of Peng et al. (2004) and our F-score on the closed track in Sighan bakeoff 2003

Sighan Bakeoff 2003	*Our F-score*	*F-score Peng et al. (2004)*
CTB	0.863	0.849
AS	0.970	0.956
HK	0.947	0.928
PK	0.953	0.941

3.2.2 Results on Sighan bakeoff 2005

Our final system achieved a F-score of 0.947 (AS), 0.943 (HK), 0.950 (PK) and 0.964 (MSR). This shows that our system successfully generalized and achieved state of the art performance on all four corpora.

Table 3 Performance of the features cumulatively, starting with the *n*-gram.

F-score	*AS*	*HK*	*PK*	*MSR*
***n*-gram**	0.943	0.946	0.950	0.961
***n*-gram (PU fixed)**		0.953		
+Unk&redupl	0.947	0.943	0.950	0.964
+Unk&redupl (PU fixed)		0.952		

Table 3 lists our results on the four corpora. We give our results using just character identity based features; character identity features plus unknown words and reduplication features. Our unknown word features only helped on AS and MSR. Both of these corpora have words that have more characters than HK and PK. This indicates that our unknown word features were more useful for corpora with segmentation standards that tend to result in longer words.

In the HK corpus, when we added in unknown word features, our performance dropped. However, we found that the testing data uses different punctuation than the training set. Our system could not distinguish new word characters from new punctuation, since having a complete punctuation list is considered external knowledge for closed track systems. If the new punctuation were not unknown to us, our performance on HK data would have gone up to 0.952 F and the unknown word features would have not hurt the system too much.

Table 4 present recalls (R), precisions (P), f-scores (F) and recalls on both unknown (R_{oov}) and known words (R_{iv}).

Table 4 Detailed performances of each corpus

	R	*P*	*F*	R_{oov}	R_{iv}
AS	0.950	0.943	0.947	0.718	0.960
HK	0.941	0.946	0.943	0.698	0.961
HK (PU-fix)	0.952	0.952	0.952	0.791	0.965
PK	0.946	0.954	0.950	0.787	0.956
MSR	0.962	0.966	0.964	0.717	0.968

3.3 Error analysis

Our system performed reasonably well on morphologically complex new words, such as 電纜**線** (CABLE in AS) and 殺人**案** (MURDER CASE in PK), where **線** (LINE) and **案** (CASE) are suffixes. However, it overgeneralized to words with frequent suffixes such as 灼**人** (it should be 灼 **人** "to burn someone" in PK) and 过**头** (it should be 回 过 **头** "to look backward" in PK). For the corpora that considered 4 character idioms as a word, our system combined most of new idioms together. This differs greatly from the results that one would likely obtain with a more traditional MaxMatch based technique, as such an algorithm would segment novel idioms.

One short coming of our system is that it is not robust enough to distinguish the difference between ordinal numbers and numbers with measure nouns. For example, 三年 (3rd year) and 三年 (three years) are not distinguishable to our system. In order to avoid this problem, it might require having more syntactic knowledge than was implicitly given in the training data.

Finally, some errors are due to inconsistencies in the gold segmentation of non-*hanzi* character. For example, "Pentium4" is a word, but "PC133" is two words. Sometimes, －8℃ is a word, but sometimes it is segmented into two words.

4 Conclusion

Our system used a conditional random field sequence model in conjunction with character identity features, morphological features and character reduplication features. We extracted our morphological information automatically to prevent overfitting Mandarin from particular Mandarin-speaking area. Our final system achieved a F-score of 0.947 (AS), 0.943 (HK), 0.950 (PK) and 0.964 (MSR).

5 Acknowledgment

Thanks to Kristina Toutanova for her generous help and to Jenny Rose Finkel who developed such a great conditional random field package. This work was funded by the Advanced Research and Development Activity's Advanced Question Answering for Intelligence Program, National Science Foundation award IIS-0325646 and a Stanford Graduate Fellowship.

References

Lafferty, John, A. McCallum, and F. Pereira. 2001. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML* 18.

Gao, Jianfeng Andi Wu, Mu Li, Chang-Ning Huang, Hongqiao Li, Xinsong Xia and Haowei Qin. 2004. Adaptive Chinese word segmentation. In *ACL-2004*.

Goh, Chooi-Ling, Masayuki Asahara, Yuji Matsumoto. 2003. Chinese unknown word identification using character-based tagging and chunking. In ACL 2003 Interactive Poster/Demo Sessions.

Ng, Hwee Tou and Jin Kiat Low. 2004. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? In *EMNLP* 9.

Peng, Fuchun, Fangfang Feng and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *COLING 2004*.

Sproat, Richard and Tom Emerson. 2003. The first international Chinese word segmentation bakeoff. In *SIGHAN 2.*

Xue, Nianwen and Libin Shen. 2003. Chinese Word Segmentation as LMR Tagging. In *SIGHAN 2.*