Pause and Stop Labeling for Chinese Sentence Boundary Detection

Hen-Hsen Huang
Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei, Taiwan
hhhuang@nlg.csie.ntu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei, Taiwan
hhchen@csie.ntu.edu.tw

Abstract

The fuzziness of Chinese sentence boundary makes discourse analysis more challenging. Moreover, many articles posted on the Internet are even lack of punctuation marks. In this paper, we collect documents written by masters as a reference corpus and propose a model to label the punctuation marks for the given text. Conditional random field (CRF) models trained with the corpus determine the correct delimiter (a comma or a full-stop) between each pair of successive clauses. Different tagging schemes and various features from different linguistic levels are explored. The results show that our segmenter achieves an accuracy of 77.48% for plain text, which is close to the human performance 81.18%. For the rich formatted text, our segmenter achieves an even better accuracy of 82.93%.

1 Introduction

To resolve sentence boundary is a fundamental issue for human language understanding. In English, sentence boundary detection (SBD) focuses on the disambiguation of the usages of punctuation marks such as period to determine if they mark the end of sentences.

In Chinese, the concept of "sentences" is fuzzier and less-defined. Native Chinese writers seldom follow the usage guidelines of punctuation marks. They often decide where to place a pause (i.e., a comma) and where to place a stop (i.e., a full-stop) in the writing according to their individual subjectivity. People tend to concatenate many clauses with commas. As a result, a Chinese sentence is often very long. That makes a text hard to be understood by both humans and machines. For example, a real world sample sentence

"這是有點霸道，但也有道理，因為他們是上市公司，每一季要向美國證管會報告總公司、附屬公司及子公司的營運及財務狀況，帳都是照一套會計原則來做，所以很多時候他們的要求，是出自一種單純的需要，而並不是故意要來欺負我們。"

could be divided into three sentences such as

"這是有點霸道，但也有道理。" 'This is a little overbearing, but is also reasonable.'

"因為他們是上市公司，每一季要向美國證管會報告總公司、附屬公司及子公司的營運及財務狀況，帳都是照一套會計原則來做。" 'Because they are listed companies and should report a summary of operation and financial status of their corporation, subsidiaries, and affiliates to the U.S. Securities quarterly, the accounts are prepared in accordance with the same set of accounting principles.'

"所以很多時候他們的要求，是出自一種單純的需要，而並不是故意要來欺負我們。" 'For this reason, their requests are usually from the simple need, not to intentionally bully us.'

The meaning from the set of shorter sentences is more concentrated and more readable than from the single longer one.

An even serious issue of Chinese punctuation marking is raised from the massive informal writing on the Internet. The articles posted frequently lack of punctuation marks. Authors usually separate clauses by whitespaces and newline symbols, and the boundaries of sentences are partially or entirely missing. Splitting an entire

Proceedings of Recent Advances in Natural Language Processing, pages 146–153,
Hissar, Bulgaria, 12-14 September 2011.

document into sentences is indispensable. For example, the following text from the Internet

“父親在一條小徑裡找到一株相思樹 正結滿了一小粒一小粒的果實
我終於知道所謂「相思果」是什麼
剪下一兩條樹枝 上面都是纍纍的紅豆
慢慢的從山上走下來
天色也跟著漸漸的黑了”

could be divided into a number of sentences with proper punctuation marks:

“父親在一條小徑裡找到一株相思樹，正結滿了一小粒一小粒的果實。” ‘My father found an acacia in a narrow path, which is covered with fruits.’

“我終於知道所謂「相思果」是什麼。” ‘I eventually knew the so called “Acacia fruit” is.’

“剪下一兩條樹枝，上面都是纍纍的紅豆。” ‘Cut a couple of branches, on which there are full of red beans.’

“慢慢的從山上走下來，天色也跟著漸漸的黑了” ‘Slowly walked down from the hill, and the sky was getting dark.’

As well, the punctuation marked text becomes more structured and more readable. At present, numerous Chinese documents on the Internet are written without the punctuation marks. To deal with those informal written data, splitting the entire document into sentences is a fundamental task as important as the Chinese word segmentation does.

In this paper, we classify the delimiter type between each pair of successive clauses into “pause” (a comma) to indicate a short stop in a sentence, and “stop” (a full-stop, an exclamation mark, or a question mark) to indicate the end of a sentence. Conditional random fields (CRFs) (Lafferty et al., 2001) are used for such a sequential labeling task. Given a text which lacks of punctuation marks or is improperly marked, the proposed model will insert or modify the punctuation marks in the text, and determine the boundaries of sentences.

The rest of this paper is organized as follows. First, we review the related work in Section 2. In Section 3, two datasets and their characteristics are presented. The labeling scheme and a variety of features are introduced in Section 4. In Section 5, the experimental results are shown and discussed. Finally, Section 6 concludes the remarks.

2 Related Work

A typical SBD task in English is to distinguish the usages of a period, including full-stop, abbreviation, number point, and a part of ellipsis (…). Various approaches are applied in this task and achieve very high performance. A rule-based model manually encoded by experts achieves an error rate of 0.9% (Aberdeen et al., 1995). The best unsupervised method achieves an error rate of 1.41% without the need of the dictionary and the abbreviation list (Mikheev, 2002). By the supervised learning approach, a modern SVM-based model achieves an even lower error rate of 0.25% (Gillick, 2009).

In Classical Chinese, there are no space and punctuation marks in the writing. As a result, all the Chinese characters in a paragraph are successive (one by one) without word, clause, and sentence boundaries. Huang et al. (2010) propose a CRF model with various features including n-gram, jump, word class, and phonetic information to segment a Classical Chinese text into clauses and achieve an F-score of 83.34%.

In Modern Chinese, Jin et al (2004) propose a method to classify the roles of commas in Chinese long sentences to improve the performance of dependency parsing. Xu et al (2005) propose a method to split a long sentence into shorter pieces to improve the performance of Chinese-English translation task. Zong and Ren (2003), and Liu and Zong (2003) segment a spoken utterance into a set of pieces. The above works focus on segmenting long sentences into shorter units for certain applications. Different from their works, recovery of the missing punctuations, and resolutions of the usages of both commas and full-stops are the major contributions of this paper.

3 Datasets

For comparison with human labeling, we sample 36 articles from Sinica corpus (Chen et al., 1996) and label them with punctuation marks by 14 native Chinese readers. Articles in this Sinica dataset are sourced from newspapers and the Internet, in which the written style and the topics are largely diverse. An article is divided into a number of fragments split by a pause punctuation (i.e., a comma) or a stop punctuation (i.e., a full-stop, an exclamation mark, or a question mark).

Dataset	#articles	#fragments	#fragments ending with a pause	#fragments ending with a stop	Average length in a fragment	#pause/#stop
Sinica dataset	36	4,498	3,175	1,323	11.76	2.40
Master dataset	1,381	296,055	204,848	91,207	10.45	2.25

Table 1. Statistics of Sinica and Master Datasets

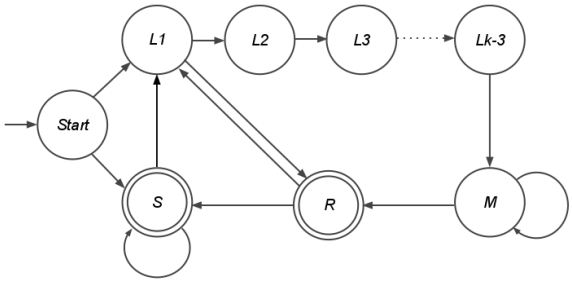

Figure 1. Markov Chain of the *k*-tag set tagging scheme

Each article is shown to three labelers without the punctuation marks, and the labelers have to label an appropriate punctuation mark at the end of each fragment. Among the 36 articles, there are 4,498 fragments in total to be labeled. The agreement between labelers is 0.554 in Fleiss' kappa, i.e., the category of moderate agreement. The mellow human agreement shows the ambiguity and the subjectivity inherent in the pause and stop labeling task.

The Sinica dataset is still not enough to be a moderate training dataset. Thus, we construct a larger Master dataset which is a collection of 1,381 articles written by Chinese masters. The masters include the modern Chinese pioneers such as Lu Xun (魯迅) and Zhu Ziqing (朱自清), the famous contemporary writers, and the professional columnists. These masters are not only the experts in Chinese writing, their writing styles are also the paradigm for Chinese learners. For this reason, the uses of punctuation marks by them can be considered as the expert-level annotation. In this way, the collection of their articles is a dataset naturally authoritative. Since the Master dataset is crawled from the Internet, the layout information like HTML tags and symbols are available in addition to the plain text. Some HTML tags such as line breaker and paragraph maker can be used as clues to sentence segmentation.

The statistics of the two datasets are shown in Table 1. The number of documents in Master dataset is 38.36 times larger than that in Sinica dataset. Besides, the number of fragments in the former dataset is 65.82 times larger than that in the latter one. The average length of a fragment in these two datasets is quite similar, i.e., 11.76 and 10.45 characters. Besides, the ratio of the number of pauses to stops is also similar, i.e., 2.40 and 2.25.

4 Labeling Method

To label the type of each delimiter between successive fragments, the sequential labeling model, CRFs, is applied. We experiment different tagging schemes and feature functions with CRF.

4.1 Tagging Scheme

The typical tagging scheme for text segmentation is 2-tag set in which two types of labels, "start" and "non-start", are used. As shown in Table 1, the ratios of the pauses to the stops are 2.40 in Sinica dataset and 2.25 in Master dataset. In other words, the classification between the class "start" and the class "non-start" is unbalanced. On average, a stop-ending clause appears after two to three pause-ending clauses.

Rather than the 2-tag set scheme, a longer tagging schemes, *k*-tag sets, are reported better in Chinese word segmentation (Xue, 2003; Zhao et al., 2006) and Classical Chinese sentence segmentation (Huang et al, 2010). We experiment different *k*-tag set schemes in pause and stop labeling. A fragment could be labeled with one of the following tags: L_1, L_2, …, L_{k-3}, *R*, *M*, and *S*.

L means *Left boundary*. The tag L_i ($1 \le i \le k-3$) labeled on fragment *f* denotes *f* is the *i*-th fragment of a sentence. The tag *R*, which means *Right boundary*, marks the last fragment of a sentence. The fragments between L_{k-3} and *R* are labeled with the tag *M* (*Middle*). A single fragment forming a sentence is labeled with the tag *S* (*Single*). The Markov Chain of the *k*-tag set tagging scheme is shown in Figure 1. For example, the fragments in the first sample in Section 1 can be labeled in the 4-tag set scheme as follows:

"這是有點霸道，" (L_1)
"但也有道理。" (*R*)
"因為他們是上市公司，" (L_1)
"每一季要向美國證管會報告總公司、附屬公司及子公司的營運及財務狀況，" (*M*)

“帳都是照一套會計原則來做。” (*R*)
“所以很多時候他們的要求，” (L_1)
“是出自一種單純的需要，” (*M*)
“而並不是故意要來欺負我們。” (*R*)

In this paper, 2-tag set, 4-tag set, and 5-tag set are explored.

4.2 Linguistic Features

Several types of features are proposed as follows.

Phonetics Level (*P*): The features include the initials, finals, and tones of the first character and the last character in a fragment. The syllabic feature useful in the speech recognition is unavailable in the written text. In this study, we use the pronunciation of each Chinese character to capture the phonetics information. In our assumption, the pronunciation combination between the last character of a fragment and the first character of the next fragment is a clue to the type (a pause or a full-stop) of successive fragments. The phonetic system is based on Mandarin Phonetic Symbols (MPS), also known as Bopomofo, in which there are 21 types of initials, 36 types of finals, and 5 types of tones.

Character Level (*C*): The features include the leftmost and the rightmost Chinese character (Hanzi) unigrams, bigrams, and trigrams of a fragment, and the number of Chinese characters in a fragment. From the empirical statistics of the distribution of Chinese words by length, 79.52% of Chinese words are covered in unigrams, bigrams, and trigrams (Chen et at., 1997).

Word Level (*W*): The features include the leftmost and the rightmost word unigrams, bigrams, and trigrams of a fragment, and the number of words in a fragment. We perform Chinese word segmentation with the Stanford Chinese word segmenter (Chang et al., 2008). As shown in Table 1, the average lengths (in Characters) of a fragment are 11.76 and 10.45 in Sinica dataset and in Master dataset, respectively. The average length of Chinese words in these two datasets is 2.49 characters. For this reason, all the characters in most fragments are able to be captured within the leftmost and the rightmost trigrams.

Part-of-Speech Level (*POS*): The features include the leftmost and the rightmost POS unigrams, bigrams, and trigrams in a fragment. Besides, the presences or absences of certain POS tags in a fragment are also checked. These tags include noun, pronoun, verb, conjunction, particle, adverb, adjective, and their combinations.

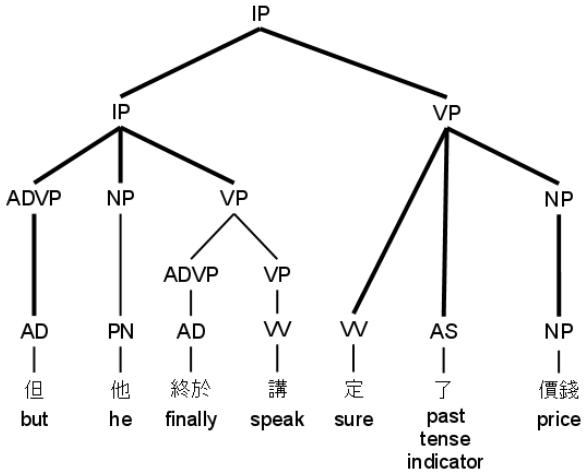

Figure 2. Extracting the top-level structure from the syntax tree

We perform POS tagging with the Stanford parser (Levy and Manning, 2003).

Syntactic Level (S): We get the syntactic tree of a fragment by the Stanford parser, and extract the structure of the upper three levels, which forms the fundamental composition of the fragment. In addition, the leftmost path and the rightmost path of the tree are also extracted. Figure 2 shows the upper three levels of the parsing tree, the leftmost path, and the rightmost path of the sample fragment in the bold edges. For instance, the structure of the upper three levels in Figure 2 formed in preorder format is IP(IP(ADVP NP VP) VP(VV AS NP)), the leftmost path is IP(IP(ADVP(AD))), and the rightmost path is IP(VP(NP(NP))).

Topic-Comment Structure (*TC*): A Chinese sentence is usually composed of a topic and several comments. The topic clause contains the topic of the sentence, and the comment clauses give more information on the topic, which is usually omitted in the comment clauses. Once a new topic appears in a clause to begin a new sentence, the sentence before the clause will be known to be complete in topic-comment structure. For example, the sentence

“我的心分外地寂寞。” ‘My heart is especially lonely.’

is a single clause and is complete in the topic-comment structure. In this example, the topic is the noun phrase “我的心” (‘My heart’), and the comment is “分外地寂寞” (‘is especially lonely’).

Consider another example:

“我從山下走下來，一路瀏覽兩旁的夜景，一路細數空中的星光。” ‘I walked from the mountain, looked at both sides of scenarios, and gazed at the stars in the sky.’

The topic is the pronoun ‘I’, and the three verb phrases, ‘Walked …’, ‘Looked at …’, and ‘Gazed at …’, are all the comments. For a given text, if one can accurately classify each fragment as a topic or a comment, the boundaries of sentences are also resolved.

To detect the topic clause is difficult. In this study, we capture the cue for topic-comment structure from the surface information. We postulate that a topic-clause tends to be a noun phrase or a complete fragment consisting of both noun phrase and verb phrase, and the comment-clause tends to be a verb phrase. For this reason, a fragment is represented in one the four types, NP, VP, NP-VP, and OTHER. In addition, the core noun in the noun phrase and the core verb in the verb phrase are also extracted.

Discourse Connective (*DC*): Some word pairs are usually used between or within sentences. We prepare a discourse connective list that contains 33 inter-sentence connectives such as “最初 ... 目前” (originally … at present) and 348 intra-sentence connectives like “不但 … 而且” (not only … but also). The two words in a pair of inter-sentence connective are collocated across sentences. For example, the pair “最初 … 目前” is almost shown in two successive sentences respectively rather than shown in the fragments which belong to a single sentence. Therefore, this is a clear cue that a stop should be inserted between inter-sentence connectives. In the other hand, the two words in a pair of intra-sentence connective are collocated within a single sentence. In this case, no stop should be inserted between them.

For each fragment, we use four features, inter-forward, inter-backward, intra-forward, and intra-backward, to capture discourse connection between it and its preceding (successive) fragment. When fragments f_i and f_j ($i<j$) contain an inter-sentence connective, the inter-forward feature of f_i and the inter-backward feature of f_j will be increased by 1. We deal with the intra-sentence connective in the similar way. That is, the corresponding intra-forward and intra-backward features will be increased accordingly. In the current implementation, the window size is set to 2.

Collocated Word (*CW*): Rather than the connectives collected from dictionaries, numerous inter and intra sentence word pairs are automatically mined from the training data as supplements to Discourse Connective, which is relatively smaller. We collect the collocations that tend to appear between inter and intra sentences from the training data, and filter them with mutual information and classification confidence.

Layout Information (*LI*): The layout information such as whitespaces, tabs, and new-lines are usually available in the text. Moreover, the articles posted on the Internet are often embedded with a lot of HTML tags and special symbols that indicate the layout styles. Those tags includes the line breaker (<*br*>), the paragraph marker (<*p*>), the span (<*span*>), the block (<*div*>), the non-breaking space (* *), and so on. The types and the occurrences of the surrounding symbols and tags form the features to represent the layout information of a fragment.

The layout information is unavailable from Sinica dataset because it is comprised of plain text.

5 Experiments

There are three parts of experiments. In the first part, we evaluate the performances of different tagging schemes with the basic features. As results, the best tagging scheme will be utilized in the following experiments. In the second part, the performances of various features and their combinations are evaluated. The best combination of the features will be adopted in the last part of experiments. In the last part, we compare the performance of our best model with those of the labelers. All the evaluation results are reported using 5-fold cross-validation.

5.1 Evaluation Metrics

All the evaluation performances are reported in terms of accuracy, precision, recall, and F-score.

Accuracy, which measures how many pauses and stops are correctly predicted, is a metric for labeling. For evaluating sentence boundary detection, we define precision as the ratio of the predicted stops between sentences which are actually stops, recall as the ratio of the stops between sentences correctly detected as stops, and F-score as the harmonic mean of precision and recall. The last punctuation mark in an article is excluded from evaluation because it is always a stop.

Tag Set	Acc.	Precision	Recall	F-Score
2-tag set	73.84%	60.25%	44.33%	51.08%
4-tag set	**77.01%**	**65.08%**	**51.59%**	**57.55%**
5-tag set	75.75%	64.68%	46.90%	54.37%

Table 2. Comparison between tagging schemes

Features	Acc.	Precision	Recall	F-Score
P	70.76%	52.53%	33.30%	40.76%
C	77.01%	65.08%	**51.59%**	57.55%
W	76.95%	66.04%	48.79%	56.12%
POS	76.77%	68.22%	43.23%	52.92%
S	71.78%	53.80%	46.56%	49.92%
TC	71.66%	55.19%	32.90%	41.22%
DC	69.73%	47.73%	2.35%	4.48%
CW	69.69%	47.58%	3.40%	6.35%
P+C+W	77.09%	65.06%	52.15%	57.90%
P+C+W+ POS	78.09%	69.08%	49.71%	57.82%
P+C+W+ POS+S	78.25%	69.02%	50.80%	**58.53%**
P+C+W+ POS+S+ TC	**78.38%**	69.63%	50.42%	58.49%
P+C+W+ POS+S+ TC+DC	77.97%	**70.76%**	46.12%	55.84%
P+C+W+ POS+S+ TC+DC+ CW	77.64%	68.99%	47.16%	56.02%
LI	78.91%	**99.97%**	30.82%	47.12%
P+C+W+ POS+S+ LI	82.74%	78.15%	**59.50%**	67.56%
P+C+W+ POS+S+ TC+LI	**82.93%**	78.90%	59.38%	**67.76%**

Table 3. Comparison among features

5.2 Tagging Scheme

The 2-tag set, 4-tag set, and 5-tag set schemes are trained over the Master dataset with the feature set on Character Level (i.e., C feature type in Section 4.2). As a result, the 4-tag set scheme outperforms the others. In the following experiments, the tag scheme is fixed to the 4-tag set.

5.3 Features

We train the model with various features over the Master dataset, and the results are listed in Table 3. The abbreviation of each feature is shown in Section 4.2. Firstly, we focus on the results when the layout information is unavailable.

Among the individual features, Character Level (*C*) features achieve the highest accuracy of 77.01% in pause and stop labeling and F-score of 57.55% in sentence boundary detection. Discourse Connective (*DC*) and Collocated Word (*CW*) suffer from the rarely matched patterns, so that the performance is out of expectation. Since the word pairs in Collocated Word are mined from the training data, we can lower the filter threshold to increase the coverage of Collocated Word. However, by adding the lower confident word pairs, the overall performance gets decreased at all.

A word is a more meaningful unit than a character in Chinese. However, the features from Word Level (*W*) are slightly inferior to those from Character Level (*C*) in our experiments. After analyzing the wrongly classified examples, we found that the Chinese word segmentation errors propagate to sentence boundary detection task. In addition, many clue words such as "了" (paste tense indicator), "嗎" (interrogative particle), and "吧" (particle used after an imperative sentence) are single character words, hence Character Level (*C*) features cover these words as well. Part-of-speech not only has the highest precision among all the single feature set, but also improves the precision when it is combined with the other features.

Although the features from Character Level (*C*) play a crucial role in the experiments, they only capture the first three and the last three characters. All of the information in the middle of fragment is missing. We try to capture that information by Syntactic Level (*S*), Topic-Comment Structure (*TC*), Discourse Connective (*DC*), and Collocated Word (*CW*). The experimental results show the combination of features on Phonetics Level (*P*), Character Level (*C*), Word Level (*W*), Part-of-Speech Level (*POS*), Syntactic Level (*S*), and Topic-Comment Structure (*TC*) achieves the best accuracy of 78.38% in pause and stop labeling and the second highest F-score of 58.49% in sentence boundary detection for the plain text. This is a significant improvement over those models trained with the features on Character or Word levels.

Layout Information *(LI)* is a special feature that achieves an extremely high precision of 99.97% and a low recall of 30.82%. The layout tags almost appear between the paragraphs or between the text blocks. In most cases, the successive clauses across two paragraphs have be inserted a full-stop. Thus, Layout Information (*LI)* is a sharp clue to roughly segment the entire article into smaller units. Combining Layout Information (*LI)* with the best models for plain text segmentation, the performance is improved by 4.55% in accuracy and 9% in F-score. Finally,

our model achieves an accuracy of 82.93% and an F-score of 67.76%.

5.4 Comparison with Human Labeling

The model trained on Master dataset is also tested on Sinica dataset to compare the performance with human labeling. Because Sinica dataset is comprised of plain text and no layout information is available, the best model for plain text is applied in this subsection.

The human performance is counted from 14 native Chinese readers' labels. The labeler who performs the best achieves an accuracy of 85.81% and an F-score of 72.15% when the author's labels are regarded as ground truth. The labeler who performs the worst has an accuracy of 77.92% and an F-score of 50.42%. The average accuracy and the F-score for all labelers are 81.18% and 67.51%, respectively. Table 4 shows the performance differences between native labelers and our model. Our model achieves 95.44% of human capability in pause and stop labeling, and 80.98% of human capability in the task of predicting sentence boundary. Overall, our model is inferior to the human average but out-perform some individuals in predicting sentence boundary.

The agreement between our model and the human labelers is 0.382 in Fleiss' kappa, and the agreements between each labeler and all the rest labelers are range from 0.363 to 0.657. This means that our model competes with native readers in this task.

Labeler	Acc.	Precision	Recall	F-score
Human Best	85.81%	70.26%	74.15%	72.15%
Huma Middle	81.15%	63.77%	72.54%	67.87%
Huma Worst	77.92%	87.97%	35.34%	50.42%
Human Average	81.18%	66.67%	68.38%	67.51%
Our Model	77.48%	65.16%	47.09%	54.67%

Table 4. Comparison between our model and the article authors

6 Conclusion

In this paper, we point out the importance of Chinese sentence boundary detection and the issue of informal writing on the Internet. To address this problem, an automatic punctuation mark label model is proposed. We test different tagging schemes and the feasibilities of various features with CRFs. For the plain text segmentation, our model with various useful linguistic features achieves accuracies of 78.38% and 77.48%, and F-scores of 58.49% and 54.67% in Master dataset and Sinica dataset, respectively. Moreover, our segmenter achieves an agreement of 0.382 compared with the human labelers. That is better than some native Chinese readers.

The best tagging scheme is 4-tag set, which outperforms the shorter and the longer tag sets in the experiments. The most useful single feature is Character (*C*), which achieves an accuracy of 77.01% and an F-score of 57.55%.

The articles ubiquitous on the Internet are usually not only plain text but embedded with layout information. For the rich formatted text, our model achieves an accuracy of 82.93% and an F-score of 67.76%. This result reveals that our model is useful to deal with the web data. Our model can be used in the application of web information extraction system, and also can be applied as the preprocessor for other tasks such as parsing and discourse boundary detection.

References

John Aberdeen, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain. 1995. MITRE: description of the Alembic system used for MUC-6. In *Proceedings of the 6th conference on Message understanding*, pages 141–155. Association for Computational Linguistics, Morristown, NJ, USA.

Pi-Chuan Chang, Michel Galley, and Chris Manning. 2008. Optimizing Chinese Word Segmentation for Machine Translation Performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, Prague, Czech Republic, June.

Aitao Chen, Jianzhang He, and Liangjie Xu. 1997. Chinese Text Retrieval Without Using a Dictionary. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42-49, Philadelphia PA, USA.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang and Hui-Li Hsu. 1996. SINICA CORPUS: Design Methodology for Balanced Corpora. In *Proceedings of PACLIC 11th Conference*, pages 167-176.

Dan Gillick. 2009. Sentence Boundary Detection and the Problem with the U.S. In Proceedings *of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, Short papers, pages 241-244, Boulder, Colorado, June. Association for Computational Linguistics.

Hen-Hsen Huang, Chuen-Tsai Sun, and Hsin-Hsi Chen. 2010. Classical Chinese Sentence Segmenta-

tion. In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 15-22, Beijing, China, August.

Mei xun Jin, Mi-Yong Kim, Dongil Kim, and JongHyeok Lee. 2004. Segmentation of Chinese Long Sentences Using Commas. In *Proceedings of SIGHAN*, pages 1-8, Barcelona, Spain.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmentation and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282-289.

Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 439-446.

Ding Liu and Chengqing Zong. 2003. Utterance Segmentation Using Combined Approach Based on Bi-directional N-gram and Maximum Entropy. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 16-23.

Andrei Mikheev. 2002. Periods, Capitalized Words, etc. *Computational Linguistics*, 28(3):289–318.

Jia Xu, Richard Zens, and Hermann Ney. 2005. Sentence Segmentation Using IBM Word Alignment Model 1. In *Proceedings of the European Association for Machine Translation (EAMT 2005)*, pages 280-287.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29-48.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2006. Effective Tag Set Selection in Chinese Word Segmentation via Conditional Random Field Modeling. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC-20)*, pages 87-94, Wuhan, China, November 1-3.

Chengqing Zong and Fuji Ren. 2003. Chinese Utterance Segmentation in Spoken Language translation. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 516-525, Mexico, Feb 16-22.