Classical Chinese Sentence Segmentation

Hen-Hsen Huang†, Chuen-Tsai Sun‡ and Hsin-Hsi Chen†
†Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
‡Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
hhhuang@nlg.csie.ntu.edu.tw ctsun@cis.nctu.edu.tw hhchen@csie.ntu.edu.tw

Abstract

Sentence segmentation is a fundamental issue in Classical Chinese language processing. To facilitate reading and processing of the raw Classical Chinese data, we propose a statistical method to split unstructured Classical Chinese text into smaller pieces such as sentences and clauses. The segmenter based on the conditional random field (CRF) model is tested under different tagging schemes and various features including n-gram, jump, word class, and phonetic information. We evaluated our method on four datasets from several eras (i.e., from the 5th century BCE to the 19th century). Our CRF segmenter achieves an F-score of 83.34% and can be applied on a variety of data from different eras.

1 Introduction

Chinese word segmentation is a well-known and widely studied problem in Chinese language processing. In Classical Chinese processing, sentence segmentation is an even more vexing issue. Unlike English and other western languages, there is no delimiter marking the end of the word in Chinese. Moreover, not only is there a lack of delimiters between the words, almost all pre-20th century Chinese is written without any punctuation marks. Figure 1 shows photocopies of printed and hand written documents from the 19th century. Within any given paragraph, the Chinese characters are printed as evenly spaced characters, with nothing to separate words from words, phrases from phrases, and sentences from sentences. Thus, inside a paragraph, explicit boundaries of sentences and clauses are lacking. In order to understand the structure, readers of Classical Chinese have to manually identify these boundaries during the reading. This process is called Classical Chinese sentence segmentation, or *Judo* (句讀).

For example, the opening lines of the Daoist classic *Zhuangzi* originally lacked segmentation:

北/north 冥/ocean 有/have 魚/fish 其/it 名/name 爲/is 鯤/Kun (a kind of big fish) 鯤/Kun 之/of 大/big 不/not 知/know 幾/how 千/thousand 里/mile 也/exclamation

The meaning of the text is hard to interpret without segmentation. Below is the identical text as segmented by a human being. It is clearly more readable.

北冥有魚/in the north ocean there is a fish
其名爲鯤 /its name is Kun
鯤之大/the size of the Kun
不知幾千里也/I don't know how many thousand miles the fish is

However, sentence segmentation in Classical Chinese is not a trivial problem. Classical Chinese sentence segmentation, like Chinese word segmentation, is inherently ambiguous. Individuals generally perform sentence segmentation in instinctive ways. To identify the boundaries of sentences and clauses, they primarily rely on their experience and sense of the language rather than on a systematic procedure. It is thus difficult to construct a set of rules or practical procedures to specify the segmentation of the infinite variety of Classical Chinese sentences.

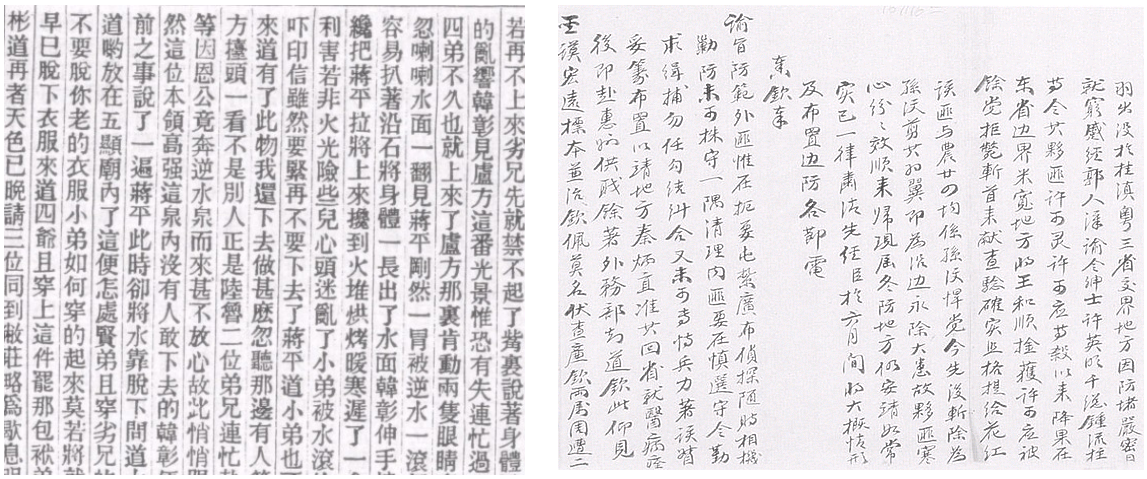

Figure 1. A Printed Page (Left) and a Hand Written Manuscript (Right) from the 19th Century.

Because of the importance of sentence segmentation, beginning in the 20th century, some editions of the Chinese classics have been labor-intensively segmented and marked with modern punctuation. However, innumerable documents in Classical Chinese from the centuries of Chinese history remain to be segmented. To aid in processing these documents, we propose an automated Classical Chinese sentence segmentation approach that enables completion of segmentation tasks quickly and accurately. To construct the sentence segmenter for Classical Chinese, the popular sequence tagging models, conditional random field (CRF) (Lafferty et al., 2001), are adopted in this study.

The rest of this paper is organized as follows. First, we describe the Classical Chinese sentence segmentation problem in Section 2. In Section 3, we review the relevant literature, including sentence boundary detection (SBD) and Chinese word segmentation. In Section 4, we introduce the tagging schemes along with the features, and show how the sentence segmentation problem can be transformed into a sequence tagging problem and decoded with CRFs. In Section 5, the experimental setup and data are described. In Section 6, we report the experimental results and discuss the properties and the challenges of the Classical Chinese sentence segmentation problem. Finally, we conclude the remarks in Section 7.

2 Problem Description

The outcomes of Classical Chinese sentence segmentation are not well-defined in linguistics at present. In general, the results of segmentation consist of sentences, clauses, and phrases. For instance, in the segmented sentence “野馬也 / 塵埃也 / 生物之以息相吹也”, “野馬也” (“the mists on the mountains like wild horses”) and “塵埃也” (“the dust in the air”) are phrases, and ” 生物之以息相吹也” (“the living creatures blow their breaths at each other”) is a clause. A sentence such as “吾以是狂而不信也” (“I do not believe it because it is ridiculous.”) is a short sentence itself, and does not require any segmentation. For a given text, there is no strict rule to determine at which level the segmentation should be performed. For instance, the opening lines of the Daoist classic *Daodejing* is “道可道非常道名可名非常名” (“The way that can be spoken is not the eternal way. The name that can be given is not the eternal name.”) which is usually segmented as “道可道 / 非常道 / 名可名 / 非常名”, but may also be segmented as “道 / 可道 / 非常道 / 名 / 可名 / 非常名”. Either segmentation is reasonable.

In this paper, we do not distinguish among the three levels of segmentation. Instead, our system learns directly from the human-segmented corpus. After training, our system will be adapted to perform human-like segmentation automatically. Further, we do not distinguish the various outcomes of Classical Chinese sentence segmentation. Instead, for the sake of convenience, every product of the segmentation process is termed “clause” in the following sections.

3 Related Work

Besides Classical Chinese, sentence boundary detection (SBD) is also an issue in English and other western languages. SBD in written texts and speech represents quite different problems. For written text, the SBD task is to distinguish periods used as the end-of-sentence indicator (full stop) from other usages, such as parts of abbreviations and decimal points. By contrast, the task of SBD in speech is closely related to the task of Classical Chinese sentence segmentation. In speech processing, the outcome of speech recognizers is a sequence of words, in which the punctuation marks are absence, and the sentence boundaries are thus lacking. To recover the syntactic structure of the original speech, SBD is required.

Like Classical Chinese sentence segmentation, the task of SBD in speech is to determine which of the inter-word boundaries in the stream of words should be marked as end-of-sentence, and then to divide the entire word sequence into individual sentences. Empirical methods are commonly employed to deal with this problem. Such methods involve many different sequence labeling models including HMMs (Shriberg et al., 2000), maximum entropy (Maxent) models (Liu et al., 2004), and CRFs (Liu et al., 2005). Among these, a CRF model used in Liu et al (2005) offered the lowest error rate.

Chinese word segmentation is a problem closely related to Classical Chinese sentence segmentation. The former identifies the boundaries of the words in a given text, while the latter identifies the boundaries of the sentences, clauses, and phrases. In contrast to sentences and clauses, the length of Chinese words is shorter, and the variety of Chinese words is more limited. Despite the minor unknown words, most of the frequent words can be handled with a dictionary predefined by Chinese language experts or extracted from the corpus automatically. However, it is impossible to maintain a dictionary of the infinite number of sentences and clauses. For these reasons, the Classical Chinese sentence segmentation problem is more challenging.

Methods of Chinese word segmentation can be mainly classified into heuristic rule-based approaches, statistical machine learning approaches, and hybrid approaches. Hybrid approaches combine the advantages of heuristic and statistical approaches to achieve better results (Gao et al., 2003; Xue, 2003; Peng et al., 2004).

Xue (2003) transformed the Chinese word segmentation problem into a tagging problem. For a given sequence of Chinese characters, the author applies a Maxent tagger to assign each character one of four positions-of-character (POC) tags, and then coverts the tagged sequence into a segmented sequence. The four POC tags used in Xue (2003) denote the positions of characters within a word. For example, the first character of a word is tagged "left boundary", the last character of a word is tagged "right boundary", the middle character of a word is tagged "middle", and a single character that forms a word by itself is tagged "single-character-word". Once the given sequence is tagged, the boundaries of words are also revealed, and the task of segmentation becomes straightforward. However, the Maxent models used in Xue (2003) suffer from an inherent the label bias problem. Peng et al (2004) uses the CRFs to address this issue. The tags used in Peng et al (2004) are of only two types, "start" and "non-start", in which the "start" tag denotes the first character of a word, and the characters in other positions are given the "non-start" tag.

The closest previous works to Classic Chinese sentence segmentation are Huang (2008) and Zhang et al. (2009). Huang combined the Xue's tagging scheme (i.e., 4-tag set) and CRFs to address the Classical Chinese sentence segmentation problem and reported an F-score of 80.96% averaged over various datasets. A similar work by Zhang et al. reported an F-score of 71.42%.

4 Methods

Conditional random field is our tagging model, and the implementation is CrfSgd 1.3[1] provided by Léon Bottou. As denoted by the tool name, the parameters in this implementation are optimized using Stochastic Gradient Descent (SGD) which convergences much faster than the common optimization algorithms such as L-BFGS and conjugate gradient (Vishwanathan, et al., 2006). To construct the sentence segmenter on

[1] http://leon.bottou.org/projects/sgd

CRF, the tagging scheme and the feature functions play the crucial roles.

4.1 Tagging Schemes

In the previous works (Huang, 2008; Zhang et al., 2009), POC tags used in Chinese word segmentation (Xue, 2003) are converted to denote the positions of characters within a clause. The 4-tag set is redefined as L ("the left boundary of a clause"), R ("the right boundary of a clause"), M ("the middle character of a clause"), and S ("a single character forming a clause"). For example, the sentence "北冥有魚其名爲鯤鯤之大不知幾千里也" should be tagged as follows.

北/L 冥/M 有/M 魚/R 其/L 名/M 爲/M 鯤/R 鯤/L 之/M 大/R 不/L 知/M 幾/M 千/M 里/M 也/R

We can easily split the sentence into clauses by making a break after each character tagged R and S and obtain the final outcome "北冥有魚 / 其名爲鯤 / 鯤之大 / 不知幾千里也".

In this work, more tagging schemes are experimented. The basic tagging scheme for segmentation is 2-tag set in which only two types of tags, "start" and "non-start", are used to label the sequence. The segmented fragments (clauses) for sentence segmentation are usually much longer than those for word segmentation. Thus, we add more middle states into the 4-tag set to model the nature of long fragments. The Markov chain of our tagging scheme is shown in Figure 2, where L2, L3, …, Lk are the additional states to extend Xue's 4-tag set. In our experiments, various k values are tested. If the k value is 1, the scheme is identical to the one used in the two previous works (Zhang et al., 2009; Huang, 2008). The 2-tag set, 4-tag set, 5-tag set and their corresponding examples are listed in Table 1. With the tagging scheme, the Classical Chinese sentence segmentation task is transformed into a sequence labeling or tagging task.

4.2 Features

Due to the flexibility of the feature function interface provided by CRFs, we apply various feature conjunctions. Besides the n-gram character patterns, the phonetic information and the part-

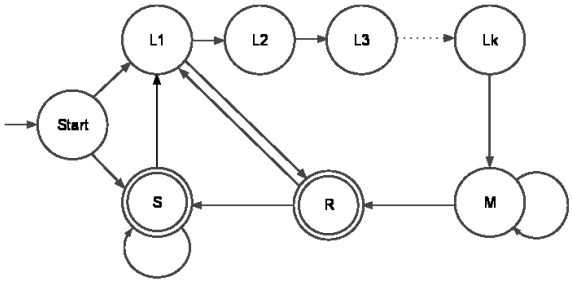

Figure 2. Markov Chain of Our Tagging Scheme.

Tag set	Tags	Example
2-tag	S: Start	**不**知其幾千里也
	N: Non-Start	不**知其幾千里也**
4-tag (k=1)	L1: Left-end	**不**知其幾千里也
	M: Middle	不**知其幾千里**也
	R: Right-end	不知其幾千里**也**
	S: Single	**性** / 猶鰥柳也
5-tag (k=2)	L1: Left-end	**不**知其幾千里也
	L2: Left-2nd	不**知**其幾千里也
	M: Middle	不知**其幾千里**也
	R: Right-end	不知其幾千里**也**
	S: Single	**性** / 猶鰥柳也

Table 1. Examples of Tag Sets.

of-speech (POS) are also included. The pronunciation of each Chinese character is labeled in three ways. The first one is Mandarin Phonetic Symbols (MPS), also known as Bopomofo, which is a phonetic system for Modern Chinese. The initial/final/tone of each character can be obtained from its MPS label.

However, Chinese pronunciation varies in the thousands of years, and the pronunciation of Modern Chinese is much different from the Classical Chinese. For this reason, two Ancient Chinese phonetic systems, Fanqie (反切) and Guangyun (廣韻), are applied to label the characters. The pronunciation of a target character is represented by two characters in the Fanqie system. The first character indicates the initial of the target character, and the second character indicates the combination of the final and the tone. The Guangyun system is in a similar manner with a smaller phonetic symbol set. There are 8,157 characters in our phonetic dictionary and the statistics are shown in Table 2.

The POS information is also considered. It is difficult to construct a Classical Chinese POS

System	#Initials	#Finals	#Tones
MPS	21	36	5
Fanqie	403		1,054
Guangyun	43		203

Table 2. Phonetic System Statistics.

POS	# Characters	Examples
Beginning	60	蓋, 唯, 雖
Middle	50	是, 或
End	45	乎, 者, 也, 矣
Interjection	20	呼, 嗟, 噫, 唉

Table 3. Four Types of POS.

tagger at this moment. Instead, we collected three types of particles that are usually placed at the beginning, at the middle, and at the end of Classical Chinese clauses. In addition, the interjections which are usually used at the end of clauses are also collected. Some examples are given in Table 3. The five feature sets and the feature templates are shown in Table 4.

5 Experiments

There are three major sets of experiments. In the 1st set of experiments, we test different tagging schemes for Classical Chinese sentence segmentation. In the 2nd set of experiments, all kinds of feature sets and their combinations are tested. The performances of the first two sets of experiments are evaluated by 10-fold cross-validation on four datasets which cross both eras and contexts. In the 3rd set of experiments, we train the system on one dataset, and test it on the others. In last part of the experiments, the generality of the datasets and the toughness of our system are tested (Peng et al., 2004). The cut-off threshold for the features is set to 2 for all the experiments. In other words, the features occur only once in the training set will be ignored. The other options of CrfSgd remain default.

5.1 Datasets

The datasets used in the evaluation are collected from the corpora of the Pre-Qin and Han Dynasties (the 5th century BCE to the 1st century BCE) and the Qing Dynasty (the 17th century CE to the 20th century CE). Chinese in the 19th century is fairly different from Chinese in the era before 0 CE. In ancient Chinese, the syntax is much simpler, the sentences are shorter, and the words are largely composed of a single character. Those are unlike later and more modern Chinese, where word segmentation is a serious issue. Given these properties, the task of segmenting

Feature Set	Template	Function
Character	$C_i, -2 \leq i \leq 2$	Unigrams
	$C_iC_{i+1}, -2 \leq i \leq 1$	Bigrams
	$C_iC_{i+1}C_{i+2}, -2 \leq i \leq 0$	Trigrams
	$C_iC_{i+2}, -2 \leq i \leq 0$	Jumps
POS	$POS\_B(C_0)$	Current character serves as a clause-beginning particle.
	$POS\_M(C_0)$	Current character serves as a clause-middle particle.
	$POS\_E(C_0)$	Current character serves as a clause-end particle.
	$POS\_I(C_0)$	Current character serves as an interjection.
MPS	$M\_I(C_0)$	The initial of current character in MPS.
	$M\_F(C_0)$	The final of current character in MPS.
	$M\_T(C_0)$	The tone of current character in MPS.
	$M\_F(C_{-1})M\_T(C_{-1})M\_I(C_0)$	The connection between successive characters.
Fanqie	$F\_I(C_0)$	The initial of current character in Fanqie.
	$F\_F(C_0)$	The final and the tone of current character in Fanqie.
	$F\_F(C_{-1})F\_I(C_0)$	The connection between successive characters.
Guangyun	$G\_I(C_0)$	The initial of the current character in Guangyun.
	$G\_F(C_0)$	The final and the tone of current character in Guangyun.
	$G\_F(C_{-1})G\_I(C_0)$	The connection between successive characters.

Table 4. Feature Templates.

Corpus	Author	Era	# of data entries	# of characters	Size of character set	Average # of characters/clause
Zuozhuan	Zuo Qiuming	500 BCE	3,381	195,983	3,238	4.145
Zhuangzi	Zhuangzi	300 BCE	1,128	65,165	2,936	5.183
Shiji	Qian Sima	100 BCE	4,778	503,890	4,788	5.049
Qing Documents	Qing Dynasty Officials	19th century	1,000	111,739	3,147	7.199

Table 5. Datasets and Statistics.

ancient Chinese sentences is easier than that of segmenting later Chinese ones. Thus, we collected texts from the pre-Qin and Han period, and from the late Qing Dynasty closer to the present, to show that our system can handle Classical Chinese as it has evolved across a span of two thousand years.

A summary of the four datasets is listed in Table 5. The *Zuozhuan* is one of earliest historical works, recording events of China in the Spring and Autumn Period (from 722 BCE to 481 BCE). The book *Zhuangzi* was named after its semi-legendary author, the Daoist philosopher Zhuangzi, who lived around the 4th century BCE. The book consists of stories and fables, in which the philosophy of the Dao is propounded. The *Shiji*, known in English as *The Records of the Grand Historian*, was written by Qian Sima in the 1st century BCE. It narrates Chinese history from 2600 BCE to 100 BCE. The *Shiji* is not only an extremely long book of more than 500,000 characters, but also the chief historical work of ancient China, exerting an enormous influence on subsequent Chinese literature and historiography.

The three ancient works are the most important classics of Chinese literature. We fetched well-segmented electronic editions of these works from the online database of the Institute of History and philology of the Academia Sinica, Taiwan.[2] Each work was partitioned into paragraphs forming a single data entry, which acted as the basic unit of training and testing. The dataset of Qing documents is selected from the Qing Palace Memorials (奏摺) related to Taiwan written in the 19th century. These documents were kindly provided by the Taiwan History Digital Library and have also been human-segmented and stored on electronic media (Chen et al., 2007). We randomly selected 1,000 paragraphs from them as our dataset.

[2]http://hanji.sinica.edu.tw

5.2 Evaluation Metrics

For Classical Chinese sentence segmentation, we define the precision P as the ratio of the boundaries of clauses which are correctly segmented to all segmented boundaries, the recall R as the ratio of correctly segmented boundaries to all reference boundaries, and the score F as the harmonic mean of precision and recall:

$$F = \frac{P \times R \times 2}{P + R}$$

Dataset	Precision	Recall	F-Score
Zuozhuan	100%	32.80%	42.73%
Zhuangzi	100%	19.84%	29.83%
Shiji	100%	14.11%	20.63%
Qing Doc.	100%	33.08%	41.42%
Average	100%	24.96%	33.65%

Table 6. Performance of Majority-Class Baseline.

Tag Set	Precision	Recall	F-Score
2-tag set	85.00%	82.16%	82.92%
4-tag set	85.11%	82.13%	82.95%
5-tag set	85.26%	82.36%	83.18%
7-tag set	84.47%	82.18%	82.74%
Baseline	100%	24.96%	33.65%

Table 7. Comparison between Tagging Schemes.

Features	Precision	Recall	F-Score
Character	85.26%	82.36%	83.18%
POS	61.04%	40.35%	43.93%
MPS	65.31%	54.00%	56.31%
Fanqie	80.96%	76.80%	77.95%
Guangyun	73.11%	69.13%	69.59%
POS + Fanqie	81.07%	74.91%	76.77%
Character + Fanqie	85.43%	82.52%	83.34%
Character + POS + Fanqie	85.67%	81.70%	82.98%

Table 8. Comparison between Feature Sets.

Dataset	Precision	Recall	F-Score
Zuozhuan	92.83%	91.56%	91.79%
Zhuangzi	81.02%	78.87%	79.34%
Shiji	80.79%	78.10%	78.99%
Qing Doc.	87.07%	81.53%	83.24%
Average	85.43%	82.52%	83.34%

Table 9. Performance on Four Datasets.

6 Results

Our baseline is a majority-class tagger which always regards the whole paragraph as a single sentence (i.e., never segments). In Table 6, the performance of the baseline is given. In the 1st set of experiments, four tagging schemes are tested while the feature set is Character. The results are shown in Table 7. In the table, each of the precision, the recall, and the F-score are averaged over the four datasets for each scheme. The results show that the CRF with the 5-tag set is superior to the 4-tag set used in previous works. However, the performance is degraded when the k is larger.

In the 2nd set of experiments, the tag scheme is fixed to the 5-tag set and a number of feature set combinations are tested. The results are shown in Table 8. The performance of MPS is significantly inferior to the other two phonetic systems. As expected, the pronunciation of Classical Chinese is much different from that of Modern Chinese, thus the Ancient Chinese phonetic systems are more suitable for this work. The Fanqie has a surprisingly performance close to the Character. However, performance of the combination of Character and Fanqie is similar to the performance of Character only model. This result indicates that the phonetic information is an important clue to Classical Chinese sentence segmentation but such information is mostly already covered by the characters. Besides, the simple POS features do not help a lot. The higher precision and the lower recall of the POS features show that the particles such as 之/乎/者/也 is indeed a clue to segmentation, but does not catch enough cases.

The best performance comes from the combination of Character and Fanqie with the 5-tag set. We use this configuration as our final tagger. The performances of our tagger for each dataset are given in Table 9. The result shows that our tagger achieves fairly good performance on the Zuozhuan segmentation, while obtaining acceptable performance overall. Because the 19th century Chinese is more complex than ancient Chinese, what we had assumed was that segmentation of the Qing documents would more difficult. However, the results indicate that our assumption does not seem to be true. Our tagger performs the sentence segmentation on the Qing documents well, even better than on the Zhuangzi and on the Shiji. The issues of longer clauses and word segmentation described earlier in this paper do not significantly affect the performance of our system.

In the last experiments, our system is trained and tested on different datasets, and the results are presented in Table 10, where the training datasets are in the rows and the test datasets are in the columns, and the F-scores of the segmentation performance are shown in the inner entries. As expected, the results of segmentation tasks across datasets are significantly poorer than the segmentation in the first two experiments.

These results indicate that our system maintains its performance on a test dataset differing from the training dataset, but the difference in written eras between the test dataset and training dataset cannot be very large. Among all datasets, Shiji is the best training dataset. As training on Shiji and testing on the two other ancient corpora Zuozhuan and Zhuangzi, the performances of our CRF segmenter are not bad.

Training Set	Testing Set				
	Zuozhuan	Zhuangzi	Shiji	Qing doc.	Average
Zuozhuan		72.04%	59.12%	38.85%	56.67%
Zhuangzi	63.70%		52.51%	42.75%	52.99%
Shiji	76.27%	75.46%		44.11%	65.28%
Qing doc.	52.68%	53.13%	42.61%		49.47%
Average	64.22%	66.88%	51.41%	41.90%	

Table 10. F-score of Segmentation cross the Datasets.

7 Conclusion

Our Classical Chinese sentence segmentation is important for many applications such as text mining, information retrieval, corpora research, and digital archiving. To aid in processing such kind of data, an automatic sentence segmentation system is proposed. Different tagging schemes and various features are introduced and tested. Our system was evaluated using three sets of experiments. Five main results are derived. First, the CRF segmenter achieves an F-score of 91.79% in the best case and 83.34% in overall performance. Second, a little longer tagging scheme improves the performance. Third, the phonetic information, especially sourced from Fanqie, is an important clue for Classical Chinese sentence segmentation and may be useful in the related tasks. Fourth, our method performs well on data from various eras. In the experiments, texts from both 500 BCE and the 19th century were well-segmented. Last, the CRF segmenter maintains a certain level of performance in situations which the test data and the training data differ in authors, genres, and written styles, but eras in which they were produced are sufficiently close.

References

Chen, Szu-Pei, Jieh Hsiang, Hsieh-Chang Tu, and Micha Wu. 2007. On Building a Full-Text Digital Library of Historical Documents. In *Proceedings of the 10th International Conference on Asian Digital Libraries, Lecture Notes in Computer Science, Springer-Verlag 4822*:49-60.

Gao, Jianfeng, Mu Li, and Chang-Ning Huang. 2003. Improved Source-Channel Models for Chinese Word Segmentation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 272-279.

Huang, Hen-Hsen. 2008. *Classical Chinese Sentence Division by Sequence Labeling Approaches*. Master's Thesis, National Chiao Tung University, Hsinchu, Taiwan.

Lafferty, John, 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*, 282-289.

Liu, Yang, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2004. Comparing and Combining Generative and Posterior Probability Models: Some Advances in Sentence Boundary Detection in Speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Liu, Yang, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using Conditional Random Fields for Sentence Boundary Detection in Speech. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 451-458. Ann Arbor, Mich., USA.

Peng, Fuchun, Fangfang Feng, and Andrew McCallum. 2004. Chinese Segmentation and New Word Detection using Conditional Random Fields. In *Proceedings of the 20th International Conference on Computational Linguistics*, 562-568.

Shriberg, Elizabeth, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. 2000. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication*, 32(1-2):127-154.

Vishwanathan, S. V. N., Nicol N. Schraudolph, Mark W. Schmidt, and Kevin P. Murphy. 2006. Accelerated training of conditional random fields with stochastic gradient methods. In *Proceedings of the 23th International Conference on Machine Learning*, 969–976. ACM Press, New York, USA.

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

Zhang, Hel, Wang Xiao-dong, Yang Jian-yu, and Zhou Wei-dong. 2009. Method of Sentence Segmentation and Punctuating for Ancient Chinese. *Application Research of Computers*, 26(9):3326-3329.