Word Segmentation for Classical Chinese Buddhist Literature

Yu-Chun Wang

Abstract

With the growth of digital humanities, information technologies take on more important roles in humanities research, including the study of religion. To analyze text for further processing, many text analysis tools treat a word as a unit. However, in Chinese, there are no word boundary markers. Word segmentation is required for processing Chinese texts. Although several word segmentation tools are available for modern Chinese, there is still no practical word segmentation tool for Classical Chinese, especially for Classical Chinese Buddhist literature. In this paper, we adopt unsupervised and supervised learning techniques to build Classical Chinese word segmentation approaches for processing Buddhist literature. Normalized variation of branching entropy (nVBE) is adopted for unsupervised word segmentation. Conditional random fields (CRF) are used to generate supervised models for Classical Chinese word segmentation. The performance of our word segmentation approach achieves an F-score of up to 0.9396. The experimental results show that our proposed method is effective for correctly segmenting most Classical Chinese sentences in Buddhist literature. Our word segmentation method can be a fundamental tool for further text analysis and processing research, such as word embedding, syntactic parsing, and semantic labeling.

1 Introduction

Word segmentation has been an essential research topic for Chinese text processing since the late 1990s. Because the Chinese writing system, unlike Western languages, does not use spaces to separate words, word segmentation is required for processing Chinese texts. With the growth of machine learning, numerous researchers have proposed algorithms or

1 Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts, Taiwan
approaches to deal with Chinese word segmentation (CWS) and achieved satisfying results. However, these approaches are mostly designed for modern Mandarin Chinese (Huang and Zhao 2007). Most of the Chinese Buddhist scriptures are written in Classical Chinese. The vocabulary and syntax of Classical and modern Chinese are very different. Thus, current Chinese word segmentation tools do not achieve good results on Classical Chinese texts. For the study of Chinese Buddhism, to deal with Chinese Buddhist texts, we need a Classical Chinese word segmentation method that is calibrated for Chinese Buddhist texts. Some current research addresses the demand for Classical Chinese word segmentation (Tsai et al. 2017).

After decades of digitization efforts, the Chinese Buddhist Electronic Text Association (CBETA) comprises the largest open-access collection of Chinese Buddhist texts. CBETA continues to digitize Chinese Buddhist scriptures and texts. We can now analyze these high-quality Chinese digital texts and use computational means to add value to the texts.

Our first goal is to provide word-segmented versions of the Chinese Buddhist texts from CBETA. Word segmentation is the necessary and fundamental basis for advanced analysis of Chinese Buddhist literature, which underlies other forms of analysis such as word embedding, dependency parsing, syntactic parsing, and semantic role labeling. Therefore, in this paper, we propose a practical approach to performing word segmentation for Classical Chinese Buddhist literature.

2 Related Work

Word segmentation has been a highly researched topic in the Chinese natural language processing community in the past two decades. Statistical approaches have dominated this domain in the past decade. The most widely adopted statistical approach is the character-based tagging method that formulates word segmentation as a sequential tagging problem. Character-based tagging was first proposed by Xue (2003). For a given sequence of Chinese characters, Xue applied a Maxent tagger to assign each character with one of four positions-of-character (POC) tags, such as “left boundary,” “right boundary,” “middle,” and “single.” Once the given sequence is tagged, the boundaries of words are also determined. Peng, Feng,
and McCallum (2004) first applied a linear-chain conditional random fields (CRF) model to Chinese word segmentation. CRF has been shown to be the optimal algorithm for sequence classification (Rosenfeld, Feldman, and Fresko 2006). This approach has been followed by many later researchers (Tseng et al. 2005; Zhao et al. 2006, 2010; Sun, Wang, and Li 2012; Zhang et al. 2013). With the growth of deep learning, instead of extracting discrete features, many researchers started to use various neural networks for automatic feature learning and discrimination. Zheng, Chen, and Xu (2013) utilized a Convolutional Neural Network (CNN) model for Chinese word segmentation to get a performance comparable to the CRF model. Chen et al. (2015) adopted a Long Short-Term Memory (LSTM) model that can keep longer-distance dependencies to achieve better results in sequence tagging.

In addition to supervised approaches, unsupervised word segmentation is also a burgeoning research topic with various methods. One of the major unsupervised methods is based on “goodness measurement.” Zhao and Kit (2008) adopted several goodness measures in unsupervised models to compare their performance. The goodness measures they used include description length gain (DLG), accessor variety (AV), and boundary entropy (BE). Wang et al. (2011) proposed an iterative model, namely ESA (“Evaluation, Selection, and Adjustment”), with a new goodness measure algorithm which uses a local maximum strategy. However, ESA requires a manually segmented training corpus to find the best values of parameters. Magistry and Sagot (2012) proposed a new model, namely normalized variation of branching entropy (nVBE), based on the variation of branching entropy proposed by Jin and Tanaka-Ishii (2006). They added normalization and Viterbi decoding to remove most of the parameters and thresholds from the model and improve performance over the previous method.

Thus, numerous approaches for word segmentation for modern Mandarin have been proposed, but Classical Chinese word segmentation has received much less attention. Qiu and Huang (2008) proposed a heuristic hybrid Classical Chinese word segmentation method based on a maximum matching algorithm with a Chinese dictionary Hanyu Da Cidian (漢語大辭典). Shi, Li, and Chen (2010) proposed a unified word segmentation and part-of-speech
tagging approach based on CRFs for pre-Qin documents. They manually annotated the *Zuo Zhuan* (左傳) to build the CRF models with predefined features. Lee and Kong (2014) and Wong and Lee (2016) describe the first attempts to tackle the word segmentation problem for Chinese Buddhist literature. Their aim is to build dependency treebanks of the Chinese Buddhist canon based on Tripiṭaka Koreana. Word segmentation is one of the preprocessing steps to construct dependency parsing trees. They adopted CRF models with predefined features based on external dictionaries. The training dataset was compiled from only four sutras with about fifty thousand characters. Although their method achieved satisfactory results, the smaller size of their dataset and the fact that it is not publicly accessible may limit usage by others. In the research described in this paper, we constructed a dataset with wider and more diverse coverage and aim to provide word-segmented Chinese Buddhist texts based on CBETA.

3 Word Segmentation Method

Below we adopt two different methods to develop a suitable word segmentation approach for Classical Chinese Buddhist literature: normalized variation of branching entropy (nVBE) and conditional random fields (CRF). nVBE is an unsupervised word segmentation method which does not require labeled data for training. The CRF method is supervised, which requires human-labeled data to build models. These two approaches have both been proven to achieve good performance on modern Mandarin word segmentation. Building on these approaches, we attempt to develop our own Classical Chinese word segmentation method.

3.1 Model 1: Unsupervised Learning Approach

nVBE is an unsupervised method derived from the method based on branching entropy (Magistry and Sagot 2012). The major idea of branching entropy is based on the hypothesis that if sequences produced by human language were random, we would expect the branching entropy of a sequence (*n*-grams in a corpus) to decrease as we increase the length of the sequence. Thus, the variation of the branching entropy (VBE) should be negative. If the
entropy of successive tokens increases, the location is at a word border. The branching entropy can be defined as follows.

Given an \( n \)-gram \( x_{0..n} = x_{0..1} \ldots x_{n-1..n} \) with a left context \( \chi_{\rightarrow} \), the right branching entropy (RBE) can be defined as

\[
h_{\rightarrow}(x_{0..n}) = - \sum_{x \in \chi_{\rightarrow}} P(x|x_{0..n}) \log P(x|x_{0..n})
\]

Also, the left branching entropy (LBE) is defined as

\[
h_{\leftarrow}(x_{0..n}) = H(\chi_{\leftarrow}|x_{0..n})
\]

where \( \chi_{\leftarrow} \) is the right context of \( x_{0..n} \).

Next, we can estimate that the variation of branching entropy (VBE) is both left and right direction as follows:

\[
\delta h_{\rightarrow}(x_{0..n}) = h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1})
\]

\[
\delta h_{\leftarrow}(x_{0..n}) = h_{\leftarrow}(x_{0..n}) - h_{\leftarrow}(x_{1..n})
\]

The VBEs are not directly comparable for strings of different lengths and need to be normalized. Following Magistry and Sagot 2012, VBEs are normalized by subtracting the mean of the VBEs of strings of the same length. Then, for different lengths of \( n \)-grams, the distributions of the VBEs at different positions inside the \( n \)-gram are compared to determine its boundaries. Therefore, for a sequence, the word segmentation problem can be formulated as a maximization problem to find the best segmentation that can generate the maximal nVBE.

For a character sequence \( s \), if we call \( \text{Seg}(s) \) the set of all the possible segmentations, then we are looking for

\[
\arg \max_{w \in \text{Seg}(s)} \sum_{w_i \in W} a(w_i) \cdot \text{len}(w_i)
\]

where \( W \) is the segmentation corresponding to the sequence of words \( w_0 \ldots w_m \), and \( \text{len}(w_i) \) is the length of a word \( w_i \) used here to allow us to compare segments which result in a different number of words. Then \( a(w_i) \) is defined as

\[
a(x) = \delta h_{\rightarrow}(x) + \tilde{\delta} h_{\rightarrow}(x)
\]
where $\tilde{\delta}$ is the normalized VBE.

### 3.2 Model 2: Supervised Learning Approach

For the supervised learning approach, we adopt the conditional random field (CRF) method as our learning model. CRFs are undirected graphical models trained to maximize a conditional probability (Lafferty, McCallum, and Pereira 2001). A linear-chain CRF with parameters $\lambda_1, \lambda_2, \ldots$ defines a conditional probability for a state sequence $y = y_1 \ldots y_T$, given that an input sequence $x = x_1 \ldots x_T$ is

$$P_\lambda(y|x) = \frac{1}{Z_x} \exp \left( \sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t) \right)$$

(6)

where $Z_x$ is the normalization factor that makes the probability of all state sequences sum to one; $f_k(y_{t-1}, y_t, x, t)$ is often a binary-valued feature function and $\lambda_k$ is its weight. The feature functions can measure any aspect of a state transition, $y_{t-1} \rightarrow y_t$, and the entire observation sequence, $x$, centered at the current time step, $t$.

#### 3.2.1 CRF Features

As shown in equation (6), to output the probability of the label sequences, CRFs sum up all the feature functions to compute the output probability. For Classical Chinese word segmentation, we define three features as follows:

1. Input character
2. Character type
3. Maximum matching with Buddhist dictionaries

The first feature is the input character itself. The second feature is the type of input characters. Input characters are categorized into three types: Chinese character, numeric character, and punctuation character. The numeric characters are all Arabic numerals. The punctuation characters are Chinese punctuation marks, such as , . ? ! 「」 , etc.

Classical Chinese Buddhist literature is mainly written in Classical Chinese. However, unlike most Classical Chinese literature, which tends to use single or double syllable words, Classical Chinese Buddhist literature contains many longer words with multiple syllables, especially transliterations from Indic languages. Therefore, we employ an array of Buddhist
dictionaries as a feature for CRF models. Each input sentence is segmented by the maximum matching algorithm (Wong and Chan 1996) with a Buddhist dictionary to get tokens. Tokens with a length less than 2 are dropped out. For each character in the input sentence, if the character is in one of the tokens, the feature value will be ‘B’ if it is the first character of the token; ‘E’ if it is the last character of the token; and ‘I’ if it is neither the first nor the last character of the token. If the character is not in any token, the feature value is ‘O’. We use the following dictionaries:

1. Fo Guang Online Dictionary (in Chinese)\(^2\) (佛光大辭典)
2. Mandarin Dictionary of Taiwan’s Ministry of Education (in Chinese)\(^3\) (教育部重編國語辭典)
3. Chinese Buddhist Encyclopedia (in Chinese)\(^4\) (中華佛典百科全書)
4. DILA Glossary\(^5\)

For each dictionary, we treat it as a distinct feature of our CRF model.

### 3.2.2 CRF Model Training

We formulate the word segmentation problem as a sequence-tagging problem which aims to tag each input character with a predefined label. The classical Buddhist texts are separated into sentences by Chinese punctuation. Then, each character in the sentences is taken as a data row for CRF model. We adopt the \{B, I, E, S\} tagging approach, which is widely used in Chinese word segmentation. The characters in a sentence are tagged as \(B\) class if it is the first character of a word or as \(I\) class if it is in a word but neither the first character nor the last character. If the character is the last character of a word, it will be tagged as \(E\) class. Otherwise, if the character is a single-character word, it will be tagged as \(S\) class.

We adopt the CRF++ open-source toolkit.\(^6\) We train our CRF models with the unigram and bigram features over the input Chinese character sequences. The features are:

---

\(^2\) Accessed May 2019, [https://www.fgs.org.tw/fgs_book/fgs_drser.aspx](https://www.fgs.org.tw/fgs_book/fgs_drser.aspx).

\(^3\) Accessed May 2019, [http://dict.revised.moe.edu.tw/cbdic/search.htm](http://dict.revised.moe.edu.tw/cbdic/search.htm).

\(^4\) Accessed May 2019, [http://buddhism.lib.ntu.edu.tw/DLMBS/search/search_detail.jsp?seq=269284](http://buddhism.lib.ntu.edu.tw/DLMBS/search/search_detail.jsp?seq=269284).

\(^5\) Accessed May 2019, [http://glossaries.dila.edu.tw](http://glossaries.dila.edu.tw).

\(^6\) Version 0.58, [https://taku910.github.io/crfpp/](https://taku910.github.io/crfpp/).
1. Unigram: $s_{-2}, s_{-1}, s_0, s_1, s_2$

2. Bigram: $s_{-1}s_0, s_0s_1$

4 Evaluation

4.1 Dataset

For the unsupervised learning approach, because it does not require annotated data, we use the whole plain text of Taishō Tripitaka from CBETA as the training corpus. By contrast, the supervised learning approach based on CRF is in need of human-annotated data for training. Therefore, we compile a dataset with segmentation annotations from two main resources, Middle Chinese Texts from Academia Sinica and a corpus from Beijing Longquan Temple. These two resources are both based on the Taishō Tripitaka text in CBETA with annotations indicating word segmentation. Table 1 shows statistics about these two resources.

|                  | Academia Sinica | Beijing Longquan Temple |
|------------------|-----------------|-------------------------|
| Sutras           | 32              | 127                     |
| Volumes          | 250             | 127                     |
| Character count  | 1.85 million    | 0.53 million            |
Since the word segmentation criteria used by these two resources are not consistent, human experts from our team at the Dharma Drum Institute of Liberal Arts (DILA) recheck and unify both to assemble one final dataset. Figure 1 shows an example of the human-annotated texts in our dataset.

### 4.2 Evaluation Metrics

To measure the performance of our methods on Classical Chinese word segmentation, we adopt the measurements of the SIGHAN Chinese Word Segmentation Bakeoff (Sproat and Emerson 2003). The measurements used by SIGHAN include precision, recall, and F-score, which are defined as follows:

\[
\text{Precision} = \frac{|\text{correct segmented words}|}{|\text{total words the system generates}|} \quad (7)
\]

\[
\text{Recall} = \frac{|\text{correct segmented words}|}{|\text{total words in the corpus}|} \quad (8)
\]

![Figure 1. Tokenized sutra text as training data](image-url)
\[ F\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \] (9)

Precision measures the ratio indicating how many words the system generates are correct; recall measures the ratio of how many actual words the system can correctly generate; and F-score is a balanced mean of both precision and recall.

### 4.3 Evaluation Results

For evaluating the supervised model, we adopt the commonly used method of tenfold cross-validation. The whole dataset is randomly partitioned into ten subsets of equal size. Each subset is sequentially used as a test dataset for evaluating the model, and the remaining nine subsets are used as a training dataset. Although the unsupervised model does not require training data, to compare its performance with the supervised model, we also used ten segments for the unsupervised model.

Table 2 shows the performance of the unsupervised model (nVBE) on the test dataset. The overall average F-score is 0.7303. Table 3 shows the performance of the supervised model (CRF) on the same test dataset. The CRF model clearly outperforms the nVBE model: its average F-score is up to 0.9396. The CRF model thus achieves better results in word segmentation for Chinese Buddhist texts.
In order to emphasize the need for and importance of the word segmentation method for

| n-fold | Precision | Recall | F-score |
|--------|-----------|--------|---------|
| 1      | 0.7574    | 0.7116 | 0.7338  |
| 2      | 0.7454    | 0.7044 | 0.7243  |
| 3      | 0.7299    | 0.7101 | 0.7199  |
| 4      | 0.7723    | 0.7186 | 0.7445  |
| 5      | 0.7617    | 0.7173 | 0.7388  |
| 6      | 0.7554    | 0.7049 | 0.7293  |
| 7      | 0.7504    | 0.7020 | 0.7254  |
| 8      | 0.7285    | 0.7036 | 0.7159  |
| 9      | 0.7546    | 0.7032 | 0.7280  |
| 10     | 0.7714    | 0.7173 | 0.7434  |
| Average| 0.7527    | 0.7093 | 0.7303  |

| n-fold | Precision | Recall | F-score |
|--------|-----------|--------|---------|
| 1      | 0.9505    | 0.9511 | 0.9508  |
| 2      | 0.9454    | 0.9552 | 0.9503  |
| 3      | 0.9414    | 0.9498 | 0.9456  |
| 4      | 0.9345    | 0.9437 | 0.9391  |
| 5      | 0.9517    | 0.9515 | 0.9516  |
| 6      | 0.9405    | 0.9486 | 0.9446  |
| 7      | 0.9314    | 0.9459 | 0.9386  |
| 8      | 0.9505    | 0.9559 | 0.9532  |
| 9      | 0.9116    | 0.9308 | 0.9211  |
| 10     | 0.8861    | 0.9172 | 0.9014  |
| Average| 0.9344    | 0.9450 | 0.9396  |
Classical Chinese, we perform another experiment. Two popular traditional Chinese word segmentation tools, the CKIP segmenter and Jieba, are used to segment the test dataset. The CKIP segmenter\(^7\) is a Chinese word segmentation tool developed by the Chinese Knowledge and Information Processing (CKIP) lab in Academia Sinica. Jieba\(^8\) is an open-source Chinese word segmentation module which builds directed acyclic graphs for all possible word combinations and then applies dynamic programming techniques to search the best segmentation based on word frequency. Jieba has the ability to take external user-defined dictionaries for segmentation. Thus, we also add Buddhist dictionaries used in CRF features to Jieba. In addition, we also apply the forward maximum matching (FMM) algorithm with dictionaries as a baseline for word segmentation. Table 4 shows the performance comparison between the CKIP segmenter, Jieba, FMM, and our two models. The results show that the general modern Chinese word segmentation tools cannot achieve good results on Classical Chinese Buddhist texts. Our supervised model is more suitable and effective for the Classical Chinese word segmentation.

### 4.4 Comparison between Model 1 and Model 2

Although model 1 takes the full texts of Taishō Tripiṭaka to build the nVBE model, the performance of nVBE is still not comparable to our model 2, even though model 2 is trained

---

\(^7\) Accessed May 2019, http://ckipsvr.iis.sinica.edu.tw.

\(^8\) Accessed May 2019, https://github.com/fxsjy/jieba.
on a smaller dataset. After examining the segmentation results of the nVBE model, we found that the nVBE model tends to combine some fixed expressions into a word. For example, the sentence “如是我聞” (Thus have I heard) is regarded as a word by the nVBE model since these four characters appear together so frequently in Chinese Buddhist texts. By contrast, if some terms, especially transliterations, share the same prefix or suffix, the nVBE model usually partitions the prefix and the following characters. For instance, “毘舍婆” (Viśvabhu) is segmented as “毘舍/婆” because there are several terms sharing the same prefix, such as “毘舍離” (Vaiśālī) and “毘舍闍” (Piśāca). The term “阿耨多羅三藐三菩提” (anuttarā-samyak-sambodhi) partitioned as “阿耨多羅/三藐三/菩提” by the nVBE model also belongs to this case since “菩提” (bodhi) is a high-frequency term in Chinese Buddhist texts.

The drawback of the unsupervised nVBE model is that it only takes the variety of the left and right context of n-grams into account. It cannot effectively be aware of differences between morphemes and actual words. The supervised model, model 2 based on CRF, can overcome this problem because it can learn to distinguish them from the annotated data.
4.5 Error Cases of Model 2

Although our model 2 achieves good performance in all three metrics, there are still some cases that cannot be correctly segmented. Table 5 shows the top 20 error cases generated by model 2.

Table 5. Top 20 Error Cases of Model 2

| Top-n | Model segmentation | Correct segmentation | Frequency |
|-------|--------------------|----------------------|-----------|
| 1     | 不/可              | 不可                  | 410       |
| 2     | 不/能              | 不能                  | 306       |
| 3     | 乃/至              | 乃至                  | 303       |
| 4     | 說/法              | 說法                  | 241       |
| 5     | 佛/說              | 佛說                  | 214       |
| 6     | 一切/眾生          | 一切眾生              | 200       |
| 7     | 僧言              | 僧言                  | 171       |
| 8     | 無/所              | 無所                  | 167       |
| 9     | 我等              | 我/等                 | 167       |
| 10    | 名/為              | 名為                  | 165       |
| 11    | 汝等              | 汝等                  | 163       |
| 12    | 不/得              | 不得                  | 161       |
| 13    | 天人              | 天/人                 | 157       |
| 14    | 所有              | 所/有                 | 146       |
| 15    | 以為              | 以/為                 | 138       |
| 16    | 以/為              | 以為                  | 134       |
| 17    | 是時              | 是時                  | 124       |
| 18    | 作/禮              | 作禮                  | 121       |
| 19    | 何/謂              | 何謂                  | 120       |
| 20    | 不可              | 不/可                 | 117       |
“乃至”，“佛說”，“說法”，and “一切眾生” (shown in table 5). To solve this problem, a consistent segmentation standard and human correction are required. With our segmentation method, these inconsistent annotations can be extracted automatically to reduce the human efforts needed for correction.

The other case is caused by a fundamental ambiguity in word segmentation. Take the thirteenth-most-common term “天人” as an example. In some sentences, it represents a single word “天人” which means “deva”; however, in some other sentences, it should be taken as two words “天/人” because it means “devas and humans.” It is difficult for a supervised learning model to distinguish these two cases without syntactic and semantic clues.

5 Conclusion

Word segmentation is required for many forms of computational analysis. In this paper, we experiment with both unsupervised and supervised approaches to implementing word segmentation for Classical Chinese Buddhist texts. We employ the normalized variation of branching entropy (nVBE) as an example of an unsupervised model and the conditional random fields (CRF) as a supervised model. To evaluate our methods, we constructed a training set from annotated corpora created by Academia Sinica and Beijing Longquan Temple, which are both based on CBETA texts. The F-scores of our unsupervised and supervised approaches average 0.7307 and 0.9396, respectively. The results show that the supervised approach based on the CRF model is more effective for segmenting words in Classical Chinese Buddhist texts and outperforms the nVBE approach, as well as the widely used modern Chinese word segmentation methods CKIP and Jieba, which can be regarded as baseline tests. In the near future, we will continue addressing the annotation inconsistency in the dataset to improve the performance of the supervised approach and apply our method to segment all the texts in CBETA. The word segmentation service and the annotated corpora will also be open access. With this word segmentation method, more advanced natural language processing techniques can be developed for application to Classical Chinese Buddhist texts.
References

Chen, Xinchi, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. “Long Short-term Memory Neural Networks for Chinese Word Segmentation.” In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015), 1197–1206. N.p.: Association for Computational Linguistics. doi:10.18653/v1/D15-1141.

Huang Chang-ning and Zhao Hai. 2007. “Chinese Word Segmentation: A Decade Review.” [In Chinese.] Journal of Chinese Information Processing 21 (3): 8–20.

Jin, Zhihui, and Kumiko Tanaka-Ishii. 2006. “Unsupervised Segmentation of Chinese Text by Use of Branching Entropy.” In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, 428–35. N.p.: Association for Computational Linguistics. doi:10.3115/1273073.1273129.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. 2001. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.” In Proceedings of the 18th International Conference on Machine Learning (ICML 2001), edited by Carla E. Brodley and Andrea Pohoreckyj Danyluk, 282–89. San Francisco: Morgan Kaufmann.

Lee, John, and Yin Hei Kong. 2016. “A Dependency Treebank of Chinese Buddhist Texts.” Digital Scholarship in the Humanities 31 (1): 140–51. First published online September 10, 2014. doi:10.1093/llc/fqu048.

Magistry, Pierre, and Benoît Sagot. 2012. “Unsupervised Word Segmentation: The Case for Mandarin Chinese.” In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, vol. 2: Short Papers, edited by Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, 383–87. Stroudsburg, PA: Association for Computational Linguistics. https://www.aclweb.org/anthology/P12-2075.pdf.

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 (COLING '04),** 562–68. Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/1220355.1220436.

Qiu Bing and Huang Fu-Huan. 2008. “Study on the Trend of Ancient Chinese Words Based on the Word Automatic Segmentation.” [In Chinese.] *Microcomputer Information* (微计算机信息) 24: 100–102.

Rozenfeld, Binyamin, Ronen Feldman, and Moshe Fresko. 2006. “A Systematic Cross-comparison of Sequence Classifiers.” In *Proceedings of the 2006 SIAM International Conference on Data Mining*, edited by Joydeep Ghosh, Diane Lambert, David Skillicorn, and Jaideep Srivastava, 564–68. Philadelphia: Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611972764.61.

Shi Min, Li Bin, and Chen Xiaohe. 2010. “CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese.” [In Chinese.] *Journal of Chinese Information Processing* (中文信息学报) 24 (2): 39–45.

Sproat, Richard, and Thomas Emerson. 2003. “The First International Chinese Word Segmentation Bakeoff.” In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, 133–43. Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/1119250.1119269.

Sun, Xu, Houfeng Wang, and Wenjie Li. 2012. “Fast Online Training with Frequency-adaptive Learning Rates for Chinese Word Segmentation and New Word Detection.” In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, edited by Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, 253–62. Stroudsburg, PA: Association for Computational Linguistics. [https://www.aclweb.org/anthology/P12-1027](https://www.aclweb.org/anthology/P12-1027).

Tsai, Richard Tzong-Han, Yu-Ting Lai, Pi-Ling Pai, Yu-Chun Wang, Sunny Hui-Min Huang, and I-Chun Fan. 2017. “WeisoEvent: A Ming-Weiso Event Analytics Tool with Named Entity Markup and Spatial-Temporal Information Linking.” In *Digital
Tseng, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. “A Conditional Random Field Word Segmenter for SIGHAN Bakeoff 2005.” In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 168–71. N.p.: Asian Federation of Natural Language Processing. https://www.aclweb.org/anthology/I05-3027.

Wang, Hanshi, Jian Zhu, Shiping Tang, and Xiaozhong Fan. 2011. “A New Unsupervised Approach to Word Segmentation.” Computational Linguistics 37 (3): 421–54. doi:10.1162/COLI_a_00058.

Wong, Pak-kwong, and Chorkin Chan. 1996. “Chinese Word Segmentation based on Maximum Matching and Word Binding Force.” In Proceedings of the 16th International Conference on Computational Linguistics, 1:200–203. Stroudsburg, PA: Association for Computational Linguistics. https://www.aclweb.org/anthology/C96-1035; doi:10.3115/992628.992665.

Wong, Tak-sum, and John Lee. 2016. “A Dependency Treebank of the Chinese Buddhist Canon.” In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), edited by Nicoletta Calzolari et al., 1679–83. N.p.: European Language Resources Association. https://www.aclweb.org/anthology/L16-1265.

Xue, Nian-wen. 2003. “Chinese Word Segmentation as Character Tagging.” Computational Linguistics and Chinese Language Processing 8 (1): 29–47. doi:10.30019/IJCLCLP.200302.0002.

Zhang, Longkai, Houfeng Wang, Xu Sun, and Mairgup Mansur. 2013. “Exploring Representations from Unlabeled Data with Co-training for Chinese Word Segmentation.” In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, edited by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, 311–21. Stroudsburg, PA:
Zhao, Hai, 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*, 87–94. [Beijing]: Tsinghua University Press.  
https://www.aclweb.org/anthology/Y06-1012.

Zhao, Hai, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2010. “A Unified Character-based Tagging Framework for Chinese Word Segmentation.” *ACM Transactions on Asian Language Information Processing (TALIP)* 9 (2): article no. 5, pp. 1–32. doi:10.1145/1781134.1781135.

Zhao, Hai, and Chunyu Kit. 2008. “An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework.” In *Proceedings of the Third International Joint Conference on Natural Language Processing*, 1:9–16. N.p.: Asian Federation of Natural Language Processing.  
https://www.aclweb.org/anthology/I08-1002.

Zheng, Xiaqing, Hanyang Chen, and Tianyu Xu. 2013. “Deep Learning for Chinese Word Segmentation and POS Tagging.” In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, edited by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, 647–57. Stroudsburg, PA: Association for Computational Linguistics.  
https://www.aclweb.org/anthology/D13-1061/.