CLASSICAL CHINESE SENTENCE SEGMENTATION AS SEQUENCE LABELING

by

Yizhou Hu

Submitted in partial fulfillment of the
requirements for Departmental Honors in
the Department of Computer Science
Texas Christian University
Fort Worth, Texas
December 19, 2014

CLASSICAL CHINESE SENTENCE SEGMENTATION AS SEQUENCE LABELING

Project Approved:

Supervising Professor: Antonio Sanchez-Aguilar, Ph.D.
Department of Computer Science

Guangyan Chen, Ph.D.
Department of Department of Modern Language Studies

Kimberly Owczarski, Ph.D.
Department of Film-TV-Digital Media

Contents

Abstract

Classical Chinese was the medium of writing in East Asia and had since extinct, leaving large amount of texts inaccessible to the general public. Expert-produced sentence segmentations are crucial to understanding classical Chinese texts. This study proposes using various NLP models to automate such segmentation as a sequence labeling problem. Results produced by automated models such as HMM, CRF, Bidirectional LSTM and similar human reproduction are all validated against expert segmentation. CRF models reach F-1 scores higher than human performance and thus are promising for potential real-life implementations. [1]

[1] The source code, complete results and sample segmented texts of this study can be found at `github.com/xlhdh/classycn`.

1 Introduction

1.1 Background

Classical Chinese is a writing system widely used over many countries in East Asia. It evolved around 5th century BCE and remained the official writing system in China, Korea, Japan and Vietnam for hundreds of years. While classical Chinese is composed of similar, if not identical, ideographic characters that make up the modern Chinese writing system, its grammar and vocabulary are vastly different from spoken languages in any of the cultures that wrote with it, including Chinese, by early 20th century.

All of the cultures that once used classical Chinese have made transitions to writing systems consistent with their spoken languages, leaving historical documents written in classical Chinese inaccessible to any individual without having to undergo extensive training. It is especially difficult for readers who do not speak Chinese to understand classical Chinese because there is less connection between their native languages and classical Chinese.

Most classical Chinese documents were produced as pure concatenation of characters, with segmentation at only paragraph level, expecting the readers to have experienced eyes to tell sentence structures through the context. Although not expected to solve the problem entirely, segmenting the text at sentence and clause levels is widely practiced, and has been of great help in overcoming the language barrier. Automating the process should further help increasing the accessibility of historical texts.

1.2 Proposal

Sentence segmentation, or 句讀(ju dou), is one technique employed by generations of readers to understand classical Chinese.

For example, *Zuo Zhuan*, a book written sometime in the last few centuries BCE starts with the following characters:

惠/Hui(posthumous name) 公/King 元/original 妃/wife of a king 孟/Meng(name) 子/Zi(name) 孟/Meng(name) 子/Zi(name) 卒/die 繼/continue 室/room, or wife 以/by 聲/Sheng(name) 子/Zi(name) 生/bearing 隱/Yin(posthumous name) 公/King

The above verse may be segmented such that it reads:

惠公元妃孟子 King Hui's original wife (was) Mengzi

孟子卒 Mengzi died

繼室以聲子 (King Hui) continued his wife by Shengzi

生隱公 bore King Yin

The text become much more readable as the relationship between the historical figures mentioned become clearer.

While much of the classical Chinese literature published today incorporate segmentation marked by human experts, large amount of texts, especially outside of China, remain un-segmented due to the amount of time and effort needed to produce segmentation.

2 Analysis

2.1 Linguistic Characteristics of Classical Chinese

Some characteristics of classical Chinese are as follows.

2.1.1 Large Character Set

Similar to modern Chinese, the number of unique characters in a classical Chinese text is usually in the thousands, and in some cases, over ten thousand. While English and other Indo-European languages do have a most likely larger vocabulary size, there exists the option to do the analysis on the character level, as the character set size is usually under 100 counting letters, spaces and marks. But for classical Chinese, one-hot vectors to represent a character will at least be 2,000-3,000 in length, which may result computing times to become prohibitively long even if such time is linear to the vector length.

2.1.2 Low Character-Word Ratio

While Chinese Word Segmentation (CWS, which segments Chinese sentences into words of one, two, three or more characters) is a prerequisite to almost any NLP work in modern Chinese, it is not essential for classical Chinese processing because the vast majority of classical Chinese words consist of only one single character. Classical Chinese can therefore be easily tokenized like English.

However, this low ratio can also result in a higher entropy per character in classical Chinese. The first order entropy of classical Chinese is 10.2 bit per character, while mordern Chinese has 9.6 bits[20]. English has around 11 bits per word[16] but only 2.62 bits per character. Higher entropy means co-occurring words are less related and we will not be able to find as much information in the context to tag the words.

2.1.3 Isolation

Classical Chinese is also an isolating language like modern Chinese, in which each word contains close to one single morpheme (rather than multiple morphemes) on average. In other words, meanings such as tenses, third person and possession are indicated by separate words instead of different affixes of one single word. (The problem, though, is that they are in many cases not indicated at all - see 2.1.4.) Therefore, stemming (splitting words into morphemes) is not necessary.

2.1.4 Ambiguity

Classical Chinese words can be extremely ambiguous in two ways. First, the exact same form of one word can serve as different parts of speech rather flexibly, sometimes even in the same clause. Second, the same word can have several completely different meanings, or be a Named Entity. It is difficult to distinguish different meanings of the same word and treating those occurrences as the same may hurt the system's performance[9].

2.2 Data Sparsity

The two datasets used in this study contain 20 million and 200 million characters respectively, but they are still of just moderate size comparing to corpora for NLP studies nowadays. The underlying fact is that the amount of available classical Chinese text on this planet is already set and no new material will be generated. While the volume can be tremendous when we use human segmenters, it is not necessarily a large set for statistical training.

2.3 Relevant Works

A computer algorithm producing accurate results will be very efficient for the segmentation task, but unfortunately, there is no set of rules that can determine these segmentations. Experienced readers segment the material they read by their knowledge and senses to the language, which is extremely difficult to reproduce with algorithms. Nevertheless, efforts have been made to automate sentence segmentation using computer algorithms.

2.3.1 N-Gram based

In 2007 [3] proposed that the possibility of a segmentation between two characters to be modeled as the number of times they appear in the training corpus with segmentation in between divided by the total number of co-occurrences. The model is then smoothed with parameters of each of the words for pairs with zero co-occurrences in the training set. Experiment on *Lun Yu* yield P=0.526, R=0.810, F-1=0.638. Huang and Hou [11] made a similar approach except that they assigned segmentation possibility 1 to every pair of words that has co-occurred in the training set with segmentation in between. They reported a recall of 0.605. Precision was not reported.

2.3.2 CRF based

Various previous work [21, 22, 10] have used the conditional random field model for the sentence segmentation task. To the author's knowledge, all have treated it as a sequence labeling problem. Different labeling schemes include 2-tag describing whether there is a segmentation point next the character and multi-tag describing how far a character is from the nearest segmentation point. Feature functions include current and neighboring characters, mutual information and pronunciation (because classical Chinese is often in rhyme).

The best results were obtained by [10] using primarily proceeding and following 1, 2 and 3-grams as features. Their work also shows that features based on pronunciation and rhyme improves performance. Their F1 varies from 0.790 to 0.918 on *Zuo Zhuan*, *Shi Ji* and *Zhuang Zi*.

3 Research Objectives

All works in this study is based on the following two hypotheses.

Hypothesis 1 All possible identities of one character in a classical Chinese paragraph forms a distribution conditioning on the identities and properties of other characters in its immediate and distant contexts within the paragraph.

Hypothesis 2 Based on the previous hypothesis, the presence or absence of a clause-level segmentation between two adjacent characters in a classical Chinese paragraph forms a binary distribution. Such distribution is conditioned on the identities and properties of these two characters and the characters in their immediate and distant context within the paragraph.

Note that both hypotheses assume that information outside the paragraph in question do not impact the contents and segmentations. While it is certainly not true, making this assumption allows us to create models of reasonable complexity. In fact, some models further restrict the source of information to the immediate context of just one or two characters adjacent to the point in question.

4 Implementation

4.1 Modeling Schema

4.1.1 Modeling Segmented Paragraphs as Labeled Sequences

Sentence segmentation is treated as a sequence labeling problem in this work, and a basic 1-tag scheme is used. In this scheme, every segmentation point marked in between characters will be attributed to the character following it or preceding it. The start and end of a sentence will be considered as segmentations. We use “START” and “END” as filler characters preceding and following the sentence. Using the sentence in the task description as an example:

START / 惠 公 元 妃 孟 子 / 孟 子 卒 / 繼 室 以 聲 子 / 生 隱 公 / END

We define two labeling schema:

To attribute a segmentation tag to the first character of each clause:

惠/S 公/N 元/N 妃/N 孟/N 子/N 孟/S 子/N 卒/N 繼/S 室/N 以/N 聲/N 子/N 生/S 隱/N 公/N END/S

To attribute a segmentation tag to the last character of each clause:

START/S 惠/N 公/N 元/N 妃/N 孟/N 子/S 孟/N 子/N 卒/S 繼/N 室/N 以/N 聲/N 子/S 生/N 隱/N 公/S

Where N denotes no segmentation and S denotes segmentation.

4.1.2 Modeling Segmentations as Characters in the Sequence

Instead of labeling, segmentation can be modeled as a special character in line with other characters. Such models are beyond the scope of this study.

4.2 Proposed Models

Let $X = (x_1, x_2, \ldots, x_n)$ be a character sequence and let $Y = (y_1, y_2, \ldots, y_n)$ be a label sequence such that x_1 has the label y_1, …, x_n has the label y_n. Each the following models can produce a possibility $P(Y|X)$ for a given tuple (X, Y). In training, a model is fed with a collection of (X, Y) tuples C, or the training set, to find a convention such that $\prod_{(X_k,Y_k)\in C} P(Y_k|X_k)$ is maximized[4].

In labeling, a model is fed with an unlabeled sequence X and it is to find a sequence $Y = (y_1, y_2, \ldots, y_n)$ such that $P(Y|X)$is maximized as per the convention determined during training.

Typically, $P(Y|X) = \prod_{i=1}^{n} P(y_i|X)$.

4.2.1 The Hidden Markov Model

The Hidden Markov Model has been widely used in sequence labeling. It assumes the next state of the sequence only depends on the previous state. It also assumes the possibility of an output label is jointly modeled with the possibility of the current token given the previous one, and the possibility of each label given the current and

Character	C3	C4
Label	N	S
Unigrams	C2, C3, C4, C5	C3, C4, C5, C6
Bigrams	C2C3, C3C4, C4C5	C3C4, C4C5, C5C6
Vectors	V(C2), V(C3), V(C4), V(C5)	V(C3), V(C4), V(C5), V(C6)

Table 1: Illustration of features used in CRF models.

previous label. The basic first-order HMM is able to recover some information in the input sequences for this particular problem and is used as a baseline model.

The experiments in this study are conducted using NLTK's [2] first order HMM package.

4.2.2 The CRF

The conditional random field model is proved to yield strong results in similar problems in recent years as discussed in the previous section. In this study it models the possibility of a label to a character given features of the character. The features used in this study are unigrams and bigrams in a 4-gram window of 2 characters before and 2 after the segmentation, and numeric word vectors of characters in the window. Table 1 illustrates the features with the following arbitrary example:

START / C1 C2 C3 / C4 C5 C6 / END

Assuming segmentations are attributed to the last character in a clause, the features for C3 are described in the table.

Three types of vectors are used: the word2vec[13] in skip-gram and CBOW, and GloVe[12]. A set of vectors of 50, 100, 300 and 500 in size are produced with each model using either the Korean or Chinese text corpus so a total of 24 different sets of vectors are used in this study. The vectors are calculated using the entire Korean or Chinese corpus regardless of the paragraph's inclusion in the training or testing set, but information about segmentation is stripped. This demonstrates the use of unsupervised pre-training. Only vectors for characters that occur at least 3 times are calculated and all other characters are assigned zero vectors.

The experiments in this study are conducted using CRF Suite[14] and a Python wrapper pyCRFSuite[15] in a stochastic gradient descent (SGD) setting.

4.2.3 The B-LSTM

The LSTM, or Long-Short Term Memory neural network[8] is a specially designed recurrent neural network (RNN) architecture used in sequence modeling problems to address the vanishing or exploding gradient problem. While a regular RNN can only take the context before the character in question into consideration[18], the bidirectional RNN model is developed to consider the whole context by feeding the sequence into the model both forwards and backwards. Graves[5] developed a bidirectional LSTM architecture and applied it to NLP problems.

Some neural network models represent the input character (or word) as a one-hot vector, and some others use various different pre-training methods such as auto-encoders to produce vector representation of the inputs. This study experiments both approaches and the vectors used are the same as the ones used in the CRF experiments. The B-LSTM in this study is implemented using Theano[1] following the definitions in [5] and [6].

4.3 Evaluation of Labeling Results

Any one of the models discussed above takes as input a sequence of unlabeled characters (usually a paragraph) and produces the same sequence of characters with each of the characters labeled as either N or S. A test set of sequences are fed into each of the models and the output labels are then compared against the original labels for each of the characters. There are several measurements to take. For each character c in the test set T (unique c is defined by its position in relation to the whole test set, not by the identity of the character it corresponds to), let l_c and l'_c be the labels associated to c in the original sequence and the output sequence respectively. $l_c, l'_c \in \{N, S\}$.

Accuracy Accuracy describe the size of the portion that is correctly labeled.

The accuracy of the outputs on test set T is defined by:

$$Accuracy = \frac{|\{c \in S | l_c = l'_c\}|}{|S|}$$

However, accuracy is a poor indicator for this imbalanced problem - the possibilities of different labels differ greatly. For example, a simple majority class labeler that labels all the entries as N achieves 0.8+ accuracy because more than 80% of the characters do not have a associate to it.

Precision and recall addresses this issue by considering the labeler's performance on only the S label.

Precision Precision is defined by the portion of the correct labels in all the characters that are labeled S in the output.

$$P = \frac{|\{c \in T | l_c = l'_c = S\}|}{|\{c \in T | l'_c = S\}|}$$

Recall Recall is defined by the portion of the correct labels in all the characters that are labeled S in the input.

$$R = \frac{|\{c \in T | l_c = l'_c = S\}|}{|\{c \in T | l_c = S\}|}$$

F-1 Score The F-1 Score is a harmonic average of precision and recall.

$$F_1 = \frac{2 \times P \times R}{P + R}$$

5 Experiment

We validate the the hypotheses using the proposed models.

5.1 Corpora

Two corpora of very different origins and styles are used in this study.

5.1.1 The Journal of the Korean Royal Secretariat

The Journal of the Korean Royal Secretariat, or *Seungjeongwon Ilgi* (hangul: 승정원일기; hanja: 承政院日記) is a record composed daily by court officials in ancient Korea. While a large portion of it was destroyed due to wars, materials containing 242 million characters in 165,000 paragraphs from 1623 to 1910 were preserved. These records were originally written in classical Chinese with shorthand because the officials had to record conversations and discussions as they were happening. Like many other classical Chinese literature, these records contain segmentation only at paragraph level. Since the 1960s, the Korean National Institution of History had transformed the shorthand to standard Chinese characters, digitized the materials, added word/clause level segmentations, marked named entities and made a searchable database online[17].

Although the material was written entirely in classical Chinese, clues of the authors' Korean identities can be spotted. For example, Korean has an SOV (subject-object-verb) structure while Chinese is SVO (subject-verb-object), and in many occasions the book preserved the Korean word order.

5.1.2 The Twenty-Four Histories

The Twenty-Four Histories was not intended as series, rather, it is a collection of Chinese histories written by various authors from last century BCE to 18th century

Corpus	Korean Royal Diary	Chinese 24 Histories
Characters	203,152,295	21,416,547
Paragraphs	1,652,528	68,040
Clauses	36,380,672	4,061,496
% of S label	17.90%	18.96%
Avg. Characters per Paragraph	123	314
Avg. Characters per Clause	5.58	5.27
Unique Characters	16,010	12,539
Structure & Style	Consistent	Varies
Authors	Non-Native Speakers	Native Speakers
Period of Composition	1623 - 1910 AD	91 BCE - 1700s AD

Table 2: Comparison of the two corpora.

CE. As a result, the works represent a large variety of writing styles and vocabulary. The corpus was crawled from WikiSource and two books (*Han Shu* and *San Guo Zhi*) were excluded due to poor quality. The remaining 22 books totaled 21.4 million characters or 68,040 paragraphs.

5.1.3 Noise

Like most other materials, both corpora include inaccurate data such as missing pages in the original document or typos during digitization. Some HTML or other forms of annotation may also have escaped the cleaning procedure. All of these are considered noise and should not severely hurt the labeling models' performances.

5.2 Quantitative Results

The same two corpora are tested separately with various configurations on the three proposed models. For each setting, both a closed test on the training set and an open test on a held out test set are performed. Training and testing cases are paragraphs randomly drawn from the corpora. All testing sets are 1/9 size of the respective training set size.

	Label Last Char		Label First Char	
Tokens	Open Test	Close Test	Open Test	Close Test
3.8E+04	40.77%	53.54%	33.69%	44.83%
3.1E+05	35.17%	40.45%	29.04%	34.28%
2.8E+06	36.81%	38.02%	30.74%	31.30%
1.9E+07	37.35%	37.21%	29.94%	30.02%

Table 3: F-1 Scores from HMM on Chinese Corpus

5.2.1 The HMM Model

The HMM model was intended only as a baseline model in this study, however various configuration was experimented to reach for its potentials. The variables include training set size, label attribution and the two different datasets.

It is clear in Figure 1 for the Korean corpus that when training size is small, the model tends to overfit and perform much better in the closed test (on training set) than in the open test (on test set). As the training size grows, performance converges and are almost identical when training size reaches 10^8 tokens.

It is surprising to see that attributing the segmentation to the last character of a clause systematically yields better results than to the first character. It is commonly perceived that both characters immediately preceding and following the segmentation are good indicators of that segmentation, but the results shows that the bigram preceding the segmentation is a better indicator than the bigram crossing the segmentation.

cite?

Finally, as it shows in Figure 3, the model performs much better on the Korean corpus than the Chinese one. It is not unexpected because the Korean one is a government record following specific writing guidelines and regulations, while the Chinese one contains a large variety of different writing styles over a course of 18 centuries.

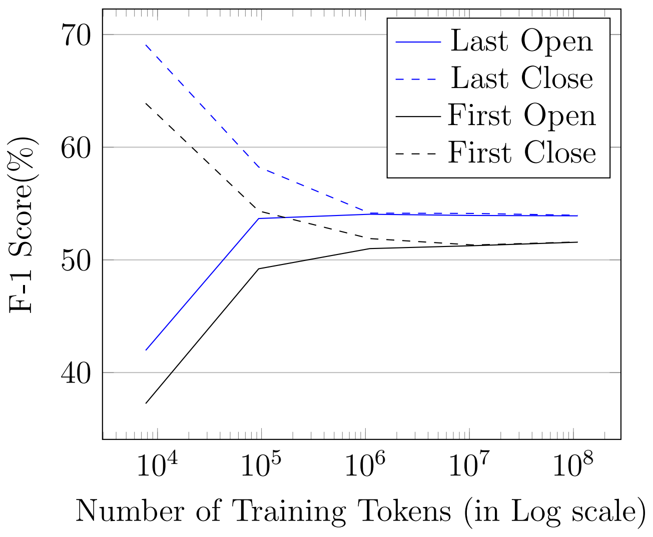

Figure 1: F-1 Scores from HMM on Korean Corpus

5.2.2 The CRF Model

The CRF experiments are performed using two types of features: discrete N-grams and continuous vector features.

The Traditional Discrete Features The CRF model yields a similar pattern of increasing test set performance and decreasing training set performance on a growing training set, but with much higher marks overall as shown in Figure 2 for the Korean corpus. Since the feature grams are the same whether a segmentation is attributed to the character preceding or following it, the difference between label attribution is almost invisible, although the scheme attributing to the preceding character still has a slight advantage.

For the Chinese corpus, however, the performance on closed testing increases as the training size grows. A possible explanation could be that the corpus contains very diverse writing styles and the model needs much more data than the Korean one to build a reasonable system at all.

Table 5 presents the average absolute value of the weights by different position of the features. We see that unigrams have mostly higher weights than bigrams, and

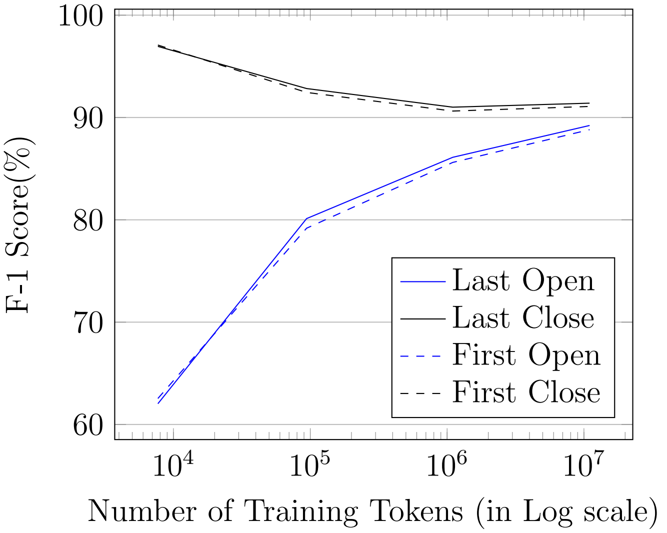

Figure 2: F-1 Scores from CRF with discrte features on Korean Corpus

	Label Last Char		Label First Char	
Tokens	Open Test	Close Test	Open Test	Close Test
3.8E+04	31.40%	53.29%	29.41%	51.78%
3.1E+05	59.54%	71.70%	58.83%	71.28%
2.8E+06	71.48%	81.06%	71.04%	80.73%
1.9E+07	78.60%	83.86%	78.26%	83.59%

Table 4: F-1 Scores from CRF with discrte features on Chinese Corpus

the two characters closest to the segmentation point are more relevant; the bigram crossing the segmentation point is less relevant than the other two.

1-gram	S-2	S-1	S	S+1	S+2
	11.46	15.33		15.13	10.98
2-gram	S-2&S-1		S-1&S+1		S+1&S+2
	10.27		8.09		11.12

Table 5: Average Absolute Value of Weights by Feature Position

The Vector Features As [19] explains, words can be represented by real-valued vectors, and such vectors can be produced using unsupervised methods, meaning the dataset does not have to be labeled.

In this study, with each corpus containing around 10,000 unique characters, the typical size of the feature bag of a word could be around 3×10^8 for each token: 10^4 for each of the 4 positions of the unigrams, and $(10^4)^2$ for each of the 3 positions of the bigrams. In practice, the number is a lot smaller at around 6×10^6 because a lot of the character pairs do not co-occur as bigrams, but the feature bag is still quite large. In the case where vector features are used instead, only 2,000 features at most are in each bag - for a vector size of 500 for each of the 4 characters in the same context window.

Table 6 shows the F-1 measurements of open tests, training with 10^6 tokens on both corpora, attributing segmentation to last character. Unfortunately, training with only the vector features yields results that are not nearly as competitive as the ones with discrete features. Using both the vector and discrete features systematically underperform the benchmark set by just discrete features, possibly because the vectors bring into the system few new information but a certain level of noise. When vector features are used alone, CBOW vectors yield the highest F-1 scores.

			Feature		
Corpus	Vector	Vector Size	Discrete	Vector	Combined
Korea	CBOW	100	86.11%	**52.81**%	85.69%
	SG			48.23%	85.40%
	GloVe			44.57%	**85.72**%
	CBOW	500		**67.81**%	85.73%
	SG			67.47%	85.63%
	GloVe			61.14%	**85.73**%
China	CBOW	100	60.66%	**29.12**%	60.91%
	SG			23.81%	**60.96**%
	GloVe			29.01%	60.95%

Table 6: F-1 Scores from CRF with Vector Features.

5.2.3 The B-LSTM Model

The results in Table 7 is not very impressive comparing with the ones obtained by CRF models with discrete features. However, for the models that are trained on the real-value vectors, they certainly overperform the CRF models with only vector features. This means that the B-LSTM has the potential of retrieving much more information than the CRF, if we can build better-designed architecture and use larger or more hidden layers.

Corpus	Tokens	Vector	Vector Size	Epoch	Open Test	Close Test
Korean	10000	One-Hot	freq=3+	0	0%	0%
	10000	CBOW	50	540	78.16%	74.83%
	10000	SG	50	446	75.24%	72.45%
	10000	GloVe	50	380	73.42%	71.10%
	10000	CBOW	500	400	80.01%	77.01%

Table 7: F-1 Scores from LSTM implementations (50 memory cells).

5.2.4 Human Segmentation Results and Acceptable Choices

Reference[7] suggests that many errors in the generated segmentation are in fact acceptable choices. Following their method, a portion of the experiment set was hand segmented by myself as an extra test set. All the models were tested on this set too. While I have been more or less exposed to the original segmentation, this also means I "learned" some original segmentation like the models did. There is no way I can remember the way a particular paragraph is segmented in the gigantic original corpus, so it is quite safe to say I am not influenced by the the correct results when segmenting. Table 8 shows the performance of the four models over the Korean corpus labeling the last character of a clause. The test set used here contains 40,000 tokens, and is disjoint with any training sets or test sets used elsewhere in the study, however, the training sets may overlap with other training sets. The HMM model was trained over 1.1E+08 tokens. The CRF model was trained over 1.1E+07 tokens using the discrete features. The B-LSTM model was trained over 1.1E+06 tokens using the CBOW vectors of size 500 as inputs. The CRF model, although trained with the smallest training set, is the best performer. It even overperforms human segmentation.

1.1E+09 LSTM instance, change appendix

Mothod	P	R	F-1	'S' Gold	'S' Yield	Train	Label
Human	89.46%	86.77%	88.10%	8141	7897	N/A	10 h
HMM	72.66%	48.05%	57.85%	8141	5384	8 min	1s
CRF	89.87%	89.58%	89.73%	8141	8115	6 min	1s
B-LSTM	85.71%	75.02%	80.01%	8141	9322	5 days	1s

Table 8: Comparison of models.

5.3 Qualitative Analysis

A short paragraph from the Korean corpus is presented here with segmentation results by each of the three models and also human. Original and segmented versions of this paragraph can be found in the appendix. The paragraph is considered representative because it contains the following components:

Concatenated Named Entities Concatenated names of people and the positions they held frequently appear in this corpus. Such concatenations can include as few as two or as many as hundreds of people. The sample paragraph reads 掌令洪光一持平洪時濟 in the beginning, where 掌令 and 持平 are position names and 洪光一 and 洪時濟 are person names. The person and the position he held are usually considered one single clause, while references to different people are separated. The CRF model and human recognized the segmentation between the two people and HMM failed to do so.

Formulated Expressions Some clauses appear repeatedly in the corpus and should be easy to predict. The last 8 characters of the paragraph 答曰勿辭亦勿退待 is a standard clause the King answers his servants and appears hundreds of times in the corpus. CRF and human labeled it correctly and HMM left out one of the three segmentation positions.

Free Expressions Free expressions are the most difficult to handle and humans depend solely on the meaning of the characters to segment them. Most of this paragraph is a free expression in which the two people discuss an issue. While

CRF and human correctly labeled most of the positions, HMM missed almost all of them.

Fixed Length Clauses Classical Chinese is often written in rhyme and/or meter, and it is desirable to have clauses of the same length. In fact, many of the clauses have exactly 4 characters. Of all the models presented here only B-LSTM and human should be sensitive to the length of the clause. , CRF also did well.

LSTM performance

6 Future Work

Several topics related to this study may worth further exploration.

Expanding the label set. In addition to the single label describing the boundary of a clause, more complicated systems may be utilized. Since the Korean corpus comes with marked named entities, they could be jointly modeled with segmentations. Models treating segmentations are special characters in line and assigns a unique label of itself (the identity label) to each character may also help with this task.

Applying the models in practice. I have not implemented the models in this study to any work that has never been segmented by human segmenters. It would be interesting analyzing the performance on real problems.

Other problems in classical Chinese processing. Translation from classical Chinese to modern Chinese, Korean, English and other languages plays a great part in introducing historic texts to modern audiences, both in their home countries and to the western world. A summary to each paragraph in gigantic corpora such as the Korean one in this study helps readers retrieve the information they need, because useful information is usually very sparse in those corpora. Many books published today in classical Chinese contain extensive hand-crafted annotations to provide the background of the author or explain certain words, etc. It is not uncommon for ⅔ of a page to be annotations and the rest to be the actual text. These processes may all benefit from automation applications.

7 Conclusion

On a corpus with consistent writing style and sufficient amount of training data, a system that quantitatively overperforms human beings can be developed to solve the classical Chinese sentence segmentation problem. The presented NLP models are able to retrieve information from the training set and produce reasonable segmentations. While HMM yields lower F-1 score and the B-LSTM requires too much training time for a similar performance, CRF with discrete features would be the choice for a practical application for both satisfactory statistics and training speed. Furthermore, this model yields very high precision marks with just the recall depending on the training size, meaning the spots it is marking as segmentation are actual stops and we just need humans to add more segmentation spots it left out. It resultes in a model suitable for pre-processing before human fine-tuning.

A Original and Segmented Paragraphs

Original /掌令洪光一/持平洪時濟啓曰/臣等之蔑識庸姿/豈有一分堪承於清朝耳目之任/而畏義分/章皇出肅/間日詣臺/未嘗論一事出一言/只呈姑停分臺/苟然充位而已/自顧慙/若無所措/果然亞憲/歷論近日臺官失職之事/至請峻其選責其任/使庸碌者/無所逃於譴何/以外他條列/無非格言/臣等已不覺瞿然自愧/而其曰/當言之事/謂屬禁令而不言/已允之啓/惟俟處分/而不論云者/尤是臣等溺職之大者/固所甘受而愧之不暇/則其何可抗顔纓/強復揚揚於臺端一步之地乎/寧被違傲之罪/不敢爲仍冒之計/累勤飭教/今始來避/所失尤大/以此情踪/其何敢一刻晏然於臺次乎/請命遞斥臣等之職/答曰/勿辭/亦勿退待/

HMM /掌令洪光一持平洪時/濟啓曰/臣等之蔑識庸姿/豈有一分堪承於清朝耳/目之任/而畏義分章皇出肅間/日詣臺未嘗論一事/出一言只呈姑停分臺苟然充位而已自顧慙/若無所措果然亞憲歷論近日臺官失職/之事/至請峻其選責/其任/使庸碌者/無所逃於譴何以外/他條列/無非格言臣等已不覺瞿然自愧而其曰/當言之事/謂屬禁令而不言已允/之啓惟俟處分而不論云者/尤是臣等溺職/之大者/固所甘受而愧/之不暇/則其何可抗顔纓強復揚揚於臺端/一步之地乎/寧被違傲/之罪不敢爲仍冒之計/累勤飭教今始來避所失尤大以此情踪/其何敢一刻晏然於臺次乎/請命遞斥臣等之職/答曰/勿辭亦勿退待/

CRF /掌令洪光一/持平洪時濟啓曰/臣等之蔑識/庸姿/豈有一分堪承於清朝耳目之任/而畏義分/章皇出肅/間日詣臺/未嘗論一事/出一言/只呈姑停分臺/苟然充位而已/自顧慙/若無所措/果然亞憲/歷論近日臺官失職之事/至請峻其選/責其任使庸碌者/無所逃於譴何以外他條列/無非格言/臣等/已不覺瞿然自愧/而其曰當言之事/謂屬禁令/而不言/已允之啓/惟俟處分/而不論云者/尤是臣等溺職之大者/固所甘受而愧之不暇/則其何可抗顔纓強復/揚揚於臺端一步之地乎/寧被違傲之罪/不敢爲仍冒之計/累勤飭教/今始來避/所失尤大/以此情踪/其何敢一刻晏然於臺次乎/請命遞斥臣等之職/答曰/勿辭/亦勿退待/

B-LSTM 掌令洪光一/持平洪時濟啓曰/臣等之蔑識庸姿/豈有一分堪/承於淸朝耳目之任/而畏義分章皇出肅間日詣臺/未嘗論一事/出一言/只呈/姑停分臺苟然充位/而已自顧慙/若無所措/果然亞憲歷論/近日臺官失職之事/至請峻其選責/其任使庸碌者/無所逃於譴/何以外/他條列/無非格言/臣等已不覺瞿然自愧/而其曰/當言之事/謂屬禁令/而不言已允之啓/惟俟處分/而不論云者/尤是臣等溺職之大者/固所甘受而愧之不暇/則其何可抗顏纓強復揚揚於臺端/一步之地乎/寧被違傲之罪/不敢爲仍冒之計/累勤飭教/今始來避所失尤大以此情踪其何敢一刻晏然於臺次乎/請命遞斥/臣等之職/答曰/勿辭亦勿退待/

Human /掌令洪光一/持平洪時濟啓曰/臣等之蔑識庸姿/豈有一分堪承於淸朝/耳目之任而畏義分/章皇出肅/間日詣臺/未嘗論一事出一言/只呈姑停分臺/苟然充位而已/自顧慙若無所措/果然亞憲歷論/近日臺官失職之事/至請峻其選/責其任使/庸碌者無所逃於譴/何以外他條列無非格言/臣等已不覺瞿然自愧/而其曰當言之事/謂屬禁令而不言/已允之啓/惟俟處分而不論云者/尤是臣等溺職之大者/固所甘受而愧之不暇/則其何可抗顏纓強/復揚揚於臺端一步之地乎/寧被違傲之罪/不敢爲仍冒之計/累勤飭教/今始來避/所失尤大/以此情踪/其何敢一刻晏然於臺次乎/請命遞斥臣等之職/答曰/勿辭/亦勿退待/

References

[1] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a cpu and gpu math compiler in python. In *Proc. 9th Python in Science Conf*, pages 1–7, 2010.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[3] Tianying Chen, Rong Chen, Lulu Pan, Hongjun Li, and Zhonghua Yu. Archaic chinese punctuating sentences based on context n-gram model. *Computer Engineering*, 33(03):192–193, 2007.

[4] Michael Collins, 2013.

[5] Alex Graves. Supervised sequence labelling with recurrent neural networks. 2012.

[6] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[7] Yuqing Guo, Haifeng Wang, and Josef Van Genabith. A linguistically inspired statistical model for chinese punctuation generation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(2):6, 2010.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[9] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational*

Linguistics: Long Papers-Volume 1, pages 873–882. Association for Computational Linguistics, 2012.

[10] Hen-Hsen Huang, Chuen-Tsai Sun, and Hsin-Hsi Chen. Classical chinese sentence segmentation. In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 15–22, 2010.

[11] Jiannian Huang and Hanqing Hou. On sentence segmentation and punctuation model for ancient books on agriculture. *Journal of Chinese Information Processing*, 22(4):31–38, 2008.

[12] Richard Socher Jeffrey Pennington and Christopher D Manning. Glove: Global vectors for word representation.

[13] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[14] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[15] Terry Peng and Mikhail Korobov. Python-crfsuite, 2007.

[16] Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[17] Byeong-Ju Shin. *Seungjeongwon Ilgi*'s value as a historical material. *Archivist*, 22:16–21, 2013.

[18] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.

[19] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th*

Annual Meeting of the Association for Computational Linguistics, pages 384–394. Association for Computational Linguistics, 2010.

[20] Dejin Wang. The probability distribution and entropy in printed chinese. *Journal of Beijing University of Aeronautics and Astronautics*, 4:010, 1988.

[21] He Zhang, Xiaodong Wang, Jianyu Yang, and Weidong Zhou. Method of sentence segmentation and punctuating based on cascaded crf. *Application Research of Computers*, (009):3326–3329, 2009.

[22] Kaixu Zhang, Yunqing Xia, and Hang Yu. Crf based approach to sentence segmentation and punctuation for ancient chinese prose. *Journal of Tsinghua University (Science & Technology)*, (10):1733–1736, 2009.