Modern Chinese Helps Archaic Chinese Processing: Finding and Exploiting the Shared Properties

Yan Song† and Fei Xia‡
†Microsoft Search Technology Center Asia
No. 5 Danling Street, Haidian District, Beijing, China
‡University of Washington
PO Box 352425, Seattle, WA 98195, USA
yansong@microsoft.com, fxia@uw.edu

Abstract

Languages change over time and ancient languages have been studied in linguistics and other related fields. A main challenge in this research area is the lack of empirical data; for instance, ancient spoken languages often leave little trace of their linguistic properties. From the perspective of natural language processing (NLP), while the NLP community has created dozens of annotated corpora, very few of them are on ancient languages. As an effort toward bridging the gap, we have created a word segmented and POS tagged corpus for Archaic Chinese using articles from *Huainanzi*, a book written during China's Western Han Dynasty (206 BC-9 AD). We then compare this corpus with the Chinese Penn Treebank (CTB), a well-known corpus for Modern Chinese, and report several interesting differences and similarities between the two corpora. Finally, we demonstrate that the CTB can be used to improve the performance of word segmenters and POS taggers for Archaic Chinese, but only through features that have similar behaviors in the two corpora.

Keywords: Archaic Chinese, HNZ Corpus, Word Segmentation, POS Tagging

1. Introduction

In the past few decades, there has been tremendous progress in the natural language processing (NLP) field, partly due to the availability of large-scale annotated corpora such as treebanks. Not surprisingly, most of the annotated corpora are for modern languages. While resources for ancient languages might not benefit many NLP applications (e.g., sentiment analysis on product reviews), they are important and valuable for linguistic research on topics such as language evolution.

The focus of this study is on Chinese, which has a long written history. According to (Wang, 1980), the history of Chinese can be generally divided into four eras: (1) Archaic era: before the Eastern Han Dynasty (around 3rd century AD); (2) Medieval era: from 3rd century till around 12th century; (3) Early Mandarin era: from Southern Song Dynasty (around 13th century) to 19th century; (4) Modern Chinese era: from the May Fourth Movement (1919 AD) to nowadays. As there is a rich collection of written documents in each era, it would be interesting to compare the characteristics of the language in different time periods and study how the language evolves throughout the time.[1]

Our work in this study consists of three parts. In the first part (Section 3.), we provide an overview of the *HNZ corpus*, a word segmented and Part-of-speech (POS) tagged corpus that we have created for Archaic Chinese using articles from *Huainanzi*. In the second part (Section 4.), we compare the HNZ corpus with the Chinese Penn Treebank (CTB), a well-known corpus for Modern Chinese, in several aspects such as word length, POS tag distributions, and word formation patterns. In the third part (Section 5.), we investigate whether NLP tools trained and tested on the HNZ corpus can benefit from the CTB, despite the differences between Archaic Chinese and Modern Chinese in general, and between the genres and content of the two corpora in particular. Note that the goal of this study is to improve the performance of NLP systems on Archaic Chinese with the help of resources for Modern Chinese; therefore, a comprehensive comparison of the two languages is out of the scope of this study.

2. Related Work

There are two types of work related to this study. The first type is about historical corpus construction. One example is the Penn Corpora of Historical English,[2] which includes the Penn-Helsinki Parsed Corpus of Middle English (Kroch and Taylor, 2000), the Penn-Helsinki Parsed Corpus of Early Modern English (Kroch et al., 2004), among others. While the corpora were developed at the University of Pennsylvania, the birth place of the English Penn Treebank (PTB) (Marcus et al., 1993), each corpus in the historical corpora has its own annotation guidelines, which are quite different from the ones for the PTB. Other studies on historical corpus construction include (Erjavec, 2012; Ogiso et al., 2012; Rögnvaldsson et al., 2012), which built corpora for ancient Slovene, Japanese and Icelandic, respectively. For Chinese, several corpora containing non-Modern Chinese text are available to the public. One of them is the Academia Sinica Tagged Corpus of Old Chinese,[3] developed by the Academia Sinica in Taiwan since 1990s. Another corpus is PKU-CCL-Corpus,[4] a large collection of unannotated text in Ancient and Modern Chinese, which was built

[1] While some researchers may treat Chinese spoken in different eras or regions as separated but related languages, others treat Chinese as one language. The discussion on this topic is out of the scope of this paper. For the sake of simplicity, we will refer to Archaic Chinese and Modern Chinese as two languages.

[2] http://www.ling.upenn.edu/histcorpora/

[3] http://old_chinese.ling.sinica.edu.tw/

[4] http://ccl.pku.edu.cn:8080/ccl_corpus

by the Center for Chinese Linguistics (CCL) at Peking University (PKU). Both corpora are designed for Chinese language research, and their main function is to provide a portal for researchers to search for examples containing certain patterns or collocations.

The second type of work is historical language processing. The idea is to build NLP systems for historical languages by taking advantage of resources for the corresponding modern languages. There are several studies on this topic (e.g., (Rayson et al., 2007; Hendrickx and Marquilhas, 2011; Scheible et al., 2012; Bollmann, 2013)), and almost all of them focus on spelling normalization; that is, they build a preprocessing module that normalizes the word spellings in the historical languages and show the spelling normalization improves system performance significantly.

However, language changes go well beyond spelling variations. For instance, Archaic Chinese is largely monosyllabic, whereas Modern Chinese has a much larger percentage of dissyllabic words (see Section 4.1.). The syntax of the two languages is also very different; for instance, a common construction in Modern Chinese, the *ba*-construction, does not exist in Archaic Chinese. Because of all these differences, a native speaker of Modern Chinese will not be able to understand Archaic Chinese articles without first taking special training on Archaic Chinese. Nevertheless, the two languages do share some similarities, such as core meanings of many Chinese characters. We propose to use features to capture similarities between an ancient language and its modern counterpart, and add those features to NLP systems to improve the system performance on the ancient language. In this study, we choose Ancient Chinese and Modern Chinese as the language pair. To evaluate our NLP systems, we first build an annotated corpus on Ancient Chinese, as explained in the next section.

3. The *HNZ* Corpus

In this section, we introduce the *HNZ* corpus[5], an Archaic Chinese corpus consisting of all the articles in the book of *Huainanzi* with word segmentation and POS tagging annotation. *Huainanzi*, also known as *Huainan Honglie*, is a collective work written by Prince Huainan, Liu An (179 BC-122 BC), and a group of his retainers in the Western Han Dynasty (206 BC-9AD). *Huainanzi* was first circulated in the Western Han Dynasty, which is near the end of the Archaic Chinese era. The book has 21 chapters, covering a wide range of topics on philosophy, astrology, geography, politics, customs, military affairs, mountains, sociology, etc. It has been described as the "Encyclopedia of the early Han Dynasty". Its abundant language capacity reveals characteristics of lexical usage in the Western Han Dynasty, and demonstrates how the usage had been transformed from the Qin Dynasty to the Han Dynasty. In this regard, *Huainanzi* contains valuable data for an in-depth analysis of Archaic Chinese. Because of these nice properties, we selected the book as the raw data for our Archaic Chinese corpus. All the manual annotation and correction was done by a Chinese linguist who is an expert on Archaic Chinese.[6]

For annotation guidelines, we start with the CTB annotation guidelines for word segmentation (Xia, 2000b) and POS tagging (Xia, 2000a), and make changes as needed. For word segmentation, we follow the same definition of wordness but the meaning of a string could change over time. For instance, the two-character string 国家 means *country* in Modern Chinese and is treated as one word, whereas it could mean *country and home* in Archaic Chinese and is therefore treated as two words under that interpretation. For POS tagging, the POS tagset for the CTB seems to be sufficient to cover the words in HNZ, so we use the same tagset for HNZ.[7]

To speed up annotation, we first manually annotated a very small portion (i.e., 750 sentences) of the HNZ corpus, and then trained a word segmenter and a POS tagger on this data. Next, we used the trained segmenter and the POS tagger to annotate the rest of the HNZ corpus, and manually corrected the system output. This semi-automatic process is much faster than starting from scratch. The whole annotated corpus consists of about 137K words. Some statistics of the corpus are presented in Section 4.

4. Modern Chinese vs. Archaic Chinese

Once the HNZ corpus has been built, we can compare it with a corpus for Modern Chinese. For the latter, we choose the Chinese Penn Treebank version 7.0 (CTB7), as it is one of the most commonly used corpora for Modern Chinese. In this section, we compare the two corpora in several aspects: word length, POS tag distributions, and word formation patterns. While some differences could be due to different genres and content of these two corpora, the comparison does shed light on some fundamental differences between Archaic Chinese and Modern Chinese.

4.1. Word Length Distribution

Some basic statistics of two corpora are given in Table 1.[8] Besides the size of the corpus, the most prominent difference between the two corpora is the average word length, which is 1.61 characters in the CTB and 1.15 characters in the HNZ.

To better understand the word length distribution, we divide the words into five categories based on their length, as in Table 2. The table shows that the large majority (85.94%) of word tokens in HNZ are monosyllabic, whereas the percentages of monosyllabic and dissyllabic words in CTB7 are about the same (50.12% vs. 42.60%); this finding is consistent with the common perception that Chinese is changing from a monosyllabic language to a language

[5] http://faculty.washington.edu/fxia/hnz/

[6] Ideally, (part of) a corpus should be annotated by multiple annotators so that inter-annotator agreement can be calculated. Unfortunately, annotating this corpus requires a lot of special training in Archaic Chinese for which we lack the resources.

[7] Some common functional words (e.g., "夫", "今夫") in the HNZ corpus are no longer used in Modern Chinese, but they can be labeled as *sentence final particle (SP)* or *other particle (MSP)*.

[8] We have used the CTB7 in several previous studies (e.g., (Song et al., 2012; Song and Xia, 2013), all of which use the same data split for training, development, and evaluation. In this study, we use only the training portion of CTB7.

Corpus	CTB7	HNZ
# of char tokens	1,409,540	158,230
# of word tokens	874,635	137,448
# of sents	38,615	7,594
Char vocabulary size	4,498	3,957
Word vocabulary size	50,035	10,847
Avg. word leng	1.61	1.15
Avg. sent leng (in chars)	36.50	20.84
Avg. sent leng (in words)	22.65	18.10

Table 1: Basic statistics of CTB7 and HNZ.

with more dissyllabic words. In both corpora, long words (length$\geq$4) are rare; they are mainly numbers in HNZ and numbers and foreign names in CTB7.

Word length	CTB7 (%)	HNZ (%)
1	438,397 (50.12)	118,119 (85.94)
2	372,567 (42.60)	18,269 (13.29)
3	46,480 (5.31)	716 (0.52)
4	9,887 (1.13)	318 (0.23)
>4	7,304 (0.84)	26 (0.02)

Table 2: Length distributions of word tokens in CTB7 and HNZ.

4.2. POS Distributions

Next, we compare the distribution of the POS tags in the two corpora. The results are in Table 3, where the tags are sorted in descending order according to their frequencies in the CTB7.

There are several observations from the table. First, the top four tags in the two corpora are the same: common noun (NN), punctuation (PU), verbs (VV), and adverb (AD). This result is not surprising given that most sentences contain a subject and/or an object (nouns), a predicate (verbs), and some adverbial modifiers (adverbs), and they end with a punctuation mark.[9]

Second, the frequencies of some POS tags in the two corpora differ a lot. For instance, out of 38 tags in the CTB7 tagset, 14 are not present in HNZ. Some of the differences are due to the genres and contents of the corpora; for instance, CTB7 has a higher percentage of numbers (CD), proper names (NR), and temporal nouns (NT) than HNZ because CTB7 contains many newswire articles which tend to have more words in those categories. Other differences are due to language change over the time. For instance, the tag *BA* is used to mark the *ba* word in the *ba*-construction. The HNZ corpus does not contain words with that tag because the *ba*-construction did not appear in Chinese until a few centuries later.[10]

[9]Punctuation did not become an integral part of Chinese written text until the 20th century, when Western punctuation marks were adopted. The punctuation marks in HNZ were not part of the original book. They were added by our annotator (after referring to other previous studies on the book) to improve the readability of the corpus.

[10]It is still up for debate whether the *ba*-construction exists in Archaic Chinese. While Mei (1990) showed some examples from the Pre-Qin Period (2100 BC-221 BC) that look like the *ba*-construction, Liu (2010) argued that those examples were not *ba*-construction and the *ba*-construction first appeared in the Six Dynasties Period (229 AD-589 AD).

POS Tag	CTB7 (%)	HNZ (%)
NN	203,398 (23.26)	35,908 (26.12)
PU	129,853 (14.85)	27,123 (19.73)
VV	124,930 (14.28)	27,974 (20.35)
AD	76,896 (8.79)	15,845 (11.53)
NR	46,168 (5.28)	2,029 (1.48)
P	29,903 (3.42)	4,637 (3.37)
CD	27,225 (3.11)	1,414 (1.03)
PN	27,061 (3.09)	4,320 (3.14)
DEG	25,896 (2.96)	4,050 (2.95)
M	23,405 (2.68)	433 (0.32)
JJ	21,020 (2.40)	602 (0.44)
DEC	19,047 (2.18)	—
NT	14,710 (1.68)	623 (0.45)
DT	14,507 (1.66)	483 (0.35)
VC	14,183 (1.62)	814 (0.59)
VA	13,245 (1.51)	1,855 (1.35)
LC	12,480 (1.43)	830 (0.60)
CC	10,072 (1.15)	758 (0.55)
AS	9,506 (1.08)	—
SP	7,331 (0.84)	4,082 (2.97)
VE	6,709 (0.77)	1,788 (1.30)
IJ	3,450 (0.39)	—
OD	2,359 (0.27)	—
MSP	2,151 (0.25)	1,518 (1.10)
CS	1,934 (0.22)	348 (0.25)
DEV	1,728 (0.20)	—
BA	1,556 (0.18)	—
ETC	1,503 (0.17)	1 (0.001)
SB	1,000 (0.11)	7 (0.005)
DER	634 (0.07)	—
LB	509 (0.06)	4 (0.003)
URL	180 (0.02)	—
NT-SHORT	27 (0.003)	—
NR-SHORT	26 (0.003)	—
ON	13 (0.001)	—
FW	8 (0.001)	—
X	6 (0.001)	—
NN-SHORT	5 (0.001)	—
Total	874,635 (100)	137,448 (100)

Table 3: The distribution of POS tags in CTB7 and HNZ. POS tags in this table are sorted in descending order according to their frequency in CTB7. Missing POS tags in HNZ are marked as "—".

4.3. Word Formation

Unlike many of the world's writing systems, the majority of characters in Chinese have meanings and can be used as words by themselves, especially in Archaic Chinese. For instance, the word "汉语" (Chinese) consists of two characters. The first character refers to the Han Dynasty or the Han people (the ethnic group that constitute approximately 92% of the population in China); the second character means *language*. So the word means the language of the Han people. "$N + N \Rightarrow N$" is just one of many common patterns for forming Chinese words. Here, the left-hand side of a pattern shows the POS tags of the characters in a word–we call these tags *cTags* (c for character)–and the right-hand side is the POS tag of the word.

We would like to compare word formation patterns in Modern and Archaic Chinese. Because cTags are not marked in the corpus, we need to infer that information somehow. A character can appear in two scenarios: (1) as a word by itself, or (2) as part of a multi-character word. In the first scenario, the cTag of the character should be the same as the POS tag of the word. In the second scenario, the cTag of each character could be different from the POS tag of the word; however, since "X + X ⇒ X" (X being a noun, a verb, etc.) is a very common word formation pattern in Chinese and we will use only the most frequent cTag of a character in the experiments in Section 5., we will simply use the POS tag of the word as the cTag of each character in that word.[11]

Char	cTags in CTB7	cTags in HNZ
上	LC:1800 VV:925 NN:450	LC:99 VV:95 NN:95
古	NR:96 NN:93 JJ:31	NT:35 NN:31 JJ:11
汉	NR:135 NN:70	NR:6 NN:2
语	NN:231 VV:12 NR:5	NN:15 VV:3

Table 4: A sample list of characters and their top three most frequent cTags in CTB7 and HNZ. The cTag list is in the form of "cTag: frequency".

Table 4 shows some examples of Chinese characters and the frequencies of their cTags in the two corpora. It turns out the cTags of many Chinese characters in the two corpora are very similar, implying that the meaning and the cTags of those characters may be relatively stable as the language evolves over time.

Once the character cTag list is obtained, we choose the most frequent cTag for each character when counting patterns that are used to form two-character words. For instance, suppose a word with the POS tag Y has two characters, c1 and c2, and the most frequent cTags for c1 and c2 are X1 and X2, respectively. We then assume the word is formed by the pattern "X1 + X2 ⇒ Y".

Table 5 lists the top ten most frequent word formation patterns in CTB7 and in HNZ. In the second and third columns, each cell has two numbers: they are the number and the percentage of dissyllabic word tokens in the corpus that are formed by this pattern. The rows are sorted by the first number in the CTB7 column. Notice that the first numbers in the HNZ column are much smaller than the numbers in the CTB7 column because only 13.29% of word tokens in HNZ are dissyllabic. The last column gives an example of dissyllabic words formed by that pattern.

The table indicates that the common word formation patterns in the two corpora are similar; for instance, seven patterns appear in both top-ten lists, and noun and verb combinations comprise the majority of two-character words in both corpora. In general, characters play an important part in composing the meaning of a multi-character word and there are certain connections between the POS tag of a word and the cTags of its characters as indicated by word formation patterns. Because the cTags of characters are relatively stable as shown in Table 4, cTag information collected from the CTB7 could help word segmentation and POS tagging of HNZ, a hypothesis which we shall test in the next section.

Pattern	CTB7 (%)	HNZ (%)	Ex
NN+NN⇒NN	104,921 (28.2)	8,347 (45.7)	政府
VV+VV⇒VV	34,908 (9.4)	725 (4.0)	教导
NR+NR⇒NR	15,443 (4.2)	90 (0.5)*	台湾
VV+NN⇒VV	14,462 (3.9)	521 (2.9)	出口
VV+NN⇒NN	11,914 (3.2)	1,022 (5.6)	知者
NR+NN⇒NN	9,656 (2.6)	108 (0.6)*	台商
NN+VV⇒VV	9,246 (2.5)	353 (1.9)	周旋
NN+NN⇒VV	8,079 (2.2)	583 (3.2)	左右
NN+VV⇒NN	6,759 (1.8)	599 (3.3)	秋分
PN+PN⇒PN	6,514 (1.8)	– *	其它
AD+AD⇒AD	6,416 (1.7)*	408 (2.2)	然而
VV+VV⇒NN	4,322 (1.2)*	268 (1.5)	要求
NN+NN⇒NR	693 (0.2)*	441 (2.4)	神农

Table 5: The top ten most common word formation patterns in the CTB7 and HNZ. The patterns are sorted by their frequencies in the CTB7. The asterisk indicates that the pattern is not among the top ten most common patterns in that corpus, but it is in the other corpus. The symbol "—" means the pattern does not occur in that corpus.

4.4. Measuring Distribution Differences

To quantitatively measure the differences between the two corpora, we built a word unigram model from each corpus and calculated the KL-divergence of the two probability distributions.[12] Since KL-divergence is asymmetric, we calculated it for both directions and reported the average. Similarly, we calculated the KL-divergence for POS unigram and cTag unigram models. The results are shown in the first column of Table 6.

For comparison, we also calculated the KL-divergence scores when the unigram models were built from different genres of CTB7. CTB7 has five genres: newswire, magazine, broadcast news, broadcast conversion, and weblog. We calculated the scores for all twenty genre pairs and reported the average on the second column of Table 6. Not surprisingly, the divergence between CTB7 and HNZ is much bigger than the one between different genres of CTB7.

	CTB7 vs. HNZ	Genres in CTB7
Word unigram	1.897	0.718
POS unigram	0.318	0.108
cTag unigram	0.406	0.081

Table 6: KL-divergence between CTB7 and HNZ and between all the genre pairs in CTB7

5. Segmentation and POS Tagging

In the previous section, we compare CTB7 and HNZ in several aspects and show the similarity and difference between

[11] We experimented with other ways of inferring cTags of characters, and all these methods yield similar results with respect to choosing the most frequent cTag for a character; therefore, we will not go into details about these alternative methods in this paper.

[12] We used add-δ for smoothing and the value of δ does not have a big effect on the KL-divergence scores. For Table 6, the δ value is set to 1.0 for all the models.

them. One interesting question is whether annotated data in one corpus could improve the performance of NLP tools on the other corpus. To test out the idea, we conduct word segmentation and POS tagging experiments on the HNZ corpus with different ways of incorporating information from CTB7.

5.1. Data

For all the experiments in this section, the test set comes from HNZ, and the training set is from HNZ, CTB7, or both depending on the settings. Because the HNZ corpus is relatively small, we ran 5-fold cross validation (1 fold for testing, 4 folds for training) and reported the average scores of five runs. Some statistics of the data split of the HNZ corpus are in Table 7. For CTB7, it is used for training only and its size is shown in Table 1.

Set	Sent	Words	Chars	Vocab
Training	6,075	109,959	126,586	9,518
Test	1,519	27,491	31,647	4,079

Table 7: Split of the HNZ corpus for training and testing the NLP systems for Archaic Chinese

5.2. Word Segmenation

We follow the general practice of treating word segmentation as a character tagging task (Xue and Shen, 2003), and build a conditional random fields (CRF) tagger. The tagset has six labels, representing a single-character word (S), the first three positions (B1, B2, B3), the last position (E), and other positions (M) of a multi-character word. We call those labels *position tags* as they represent the position of a character in a word. For instance, if "c1 c2 c3" is a word, the corresponding character-tag sequence will be "c1/B1 c2/B2 c3/E".

Table 8 lists the features used by the segmenter, where the subscript -1, 0, and +1 refer to the previous, current and next character, respectively. The first two types are character unigrams and bigrams. The *cTag* feature refers to the most frequent cTag of the current character. The process of inferring cTag was described in Section 4.3., and the cTag frequencies are collected from the training data only.

The DLG features are based on the idea of using description length gain (DLG) (Kit and Wilks, 1999) for unsupervised word segmentation (Kit, 2000; Kit, 2005). Intuitively, the DLG score of a string indicates the reduction of description length of a corpus when the string is treated as a word; the more frequent a string is in corpus and the longer the string is, the higher its DLG score is. The DLG features D_0^k basically represent the decisions made by the DLG-based unsupervised word segmenter when the current (presented by the subscript 0) character is considered to be part of a *k*-character word in the current sentence. More detail of the DLG features is available in (Song and Xia, 2012). Because DLG scores rely on only unlabeled data where the word boundary is ignored, when HNZ is used as the corpus for training, we use the whole unlabeled HNZ corpus, which includes the test data, to collect DLG scores.

We ran three sets of experiments and the results are in Table 9. The first set uses basic features only (i.e., character unigrams and bigrams). The training data come from HNZ only (Expt #1), the CTB7 only (Expt #2), or the union of the two (Expt #3). The experiments show that word segmenter trained on the CTB7 only performs poorly on the HNZ test data (#2), and adding CTB7 to HNZ did not help either (#3 vs. #1). This is due to different characteristics of Archaic and Modern Chinese. As illustrated in Table 2, 85.94% of words in HNZ are monosyllabic; therefore, most characters in HNZ should get a *S* position tag (*S* for a single-character word). In contrast, about half of words in CTB7 have multiple characters, which means that most characters in CTB7 should get a position tag other than *S*. As a result, using CTB7 to collect basic features will only hurt the performance.

Description	Features
Char Unigrams	C_{-1}, C_0, C_{+1}
Char Bigrams	$C_{-1}C_0, C_0C_{+1}, C_{-1}C_{+1}$
cTag Feature	T_0
DLG Features	$D_0^1, D_0^2, D_0^3, D_0^4, D_0^5$

Table 8: Feature templates of our CRF segmenter.

In the second set of experiments, basic features are extracted from the HNZ training data only; cTag features are collected from HNZ only (#4), CTB7 only (#5), or the union of the two corpora (#6). Expt #7 uses both corpora as well, but cTag frequency is collected from each corpus separately and there will be two copies of the cTag features, one for CTB7, the other for HNZ, following the feature augmentation idea proposed by Daume (2007). Interestingly, using cTags collected from CTB only (#5) provides a modest gain over the baseline (#1), but using cTags from HNZ (#4) only or from both corpora (#6-#7) does not help. This might be due to the fact that HNZ is small and thus cTags extracted from HNZ is less reliable than the cTags collected from CTB7.

In the third set of experiments, we add DLG features on top of basic and cTag features, where DLG scores are collected from HNZ (#8), CTB (#9), the union of two corpora (#10), or each corpus separately (#11). The results confirmed that DLG features further improve system performance. Among these experiments, HNZ (#8) works the best because adding DLG features would make the segmenter to favor treating a string appearing multiple times in a corpus as a word. When CTB7 is used, some of such strings (such as 国家) would be labeled as a word–a correct decision (the word means *country*) in Modern Chinese, but an error (the string means *country and home*) in Archaic Chinese.

While the gain achieved from adding cTag and DLG features may seem modest, it is much more significant when the amount of the training data is smaller. Figure 1 shows the F-scores of Expt #1, #5, and #8 when only x% of the training data from HNZ is used to collect basic features, where x ranges from 10 to 100. The smaller x is, the bigger gain cTag and DLG features provide.

5.3. POS Tagging

For POS tagging, we use a CRF tagger similar to the one used for word segmentation. Table 10 shows the feature

Expt id	Basic Features (BF)		Additional Features				F-score
	BF(HNZ)	BF(CTB)	cTag(HNZ)	cTag(CTB)	DLG(HNZ)	DLG (CTB)	
#1	+						<u>92.83</u>
#2		+					62.26
#3	+(u)	+(u)					89.66
#4	+		+				92.59
#5	+			+			92.92
#6	+		+(u)	+(u)			92.73
#7	+		+(f)	+(f)			92.77
#8	+			+	+		**93.25**
#9	+			+		+	93.06
#10	+			+	+(u)	+(u)	93.22
#11	+			+	+(f)	+(f)	93.17

Table 9: Performance of the word segmenter with different features and training data. A "+" symbol denotes that the feature or data set indicated by the column label is used for training. The "(u)" after the plus symbol means that the cTags are collected from the union of CTB7 and HNZ, whereas "(f)" means the cTags are collected from the two corpora separately and they are treated as two separate features. The last column shows the F-scores of word segmentation. The score of the baseline system is underlined, and the best result is in boldface.

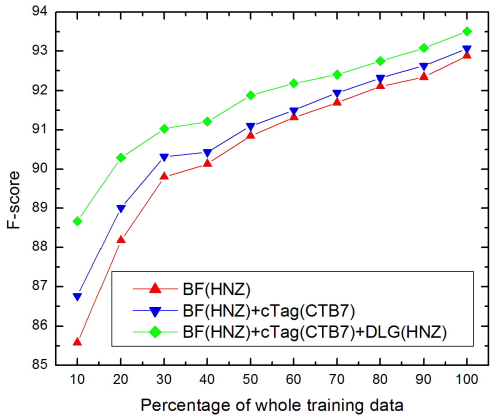

Figure 1: Performance of the word segmenter with three feature sets. Basic features (BF) come from a reduced amount of training data in HNZ; cTag features are collected from CTB7; DLG features are extracted from the whole unlabeled HNZ corpus, including the test data.

set used by the tagger, where subscript -1, 0, and +1 refer to the previous, current and next word, respectively. For word affix features, P_0 (P for prefix) and S_0 (S for suffix) refer to the first and the last character of the current word, respectively; affix features are always extracted from the same training data as word ngram features. TP_0 and TS_0 are the most frequent cTag of P_0 and S_0 in the training data, respectively, and the cTag frequencies can be extracted from either corpus or from both.

Description	Features
Word unigrams	W_{-1}, W_0, W_{+1}
Word bigrams	$W_{-1}W_0, W_0W_{+1}, W_{-1}W_{+1}$
Word affix	P_0, S_0
cTag of affix	TP_0, TS_0

Table 10: The feature set for our CRF POS tagger

Like word segmentation, we ran three sets of experiments for POS tagging, and the results are in Table 11. The table shows similar patterns as what we have observed from word segmentation experiments. First, using CTB with basic features (i.e., word ngrams) hurts the performance (e.g., Expt #2 and #3 vs. #1). Second, adding cTag features collected from CTB7 provides a boost over the baseline (#5 vs. #1).[13] Third, adding affix features further improves the performance (#8-11 vs. #4-7).

Figure 2 shows the performance of Expt #1, #5 and #8 when the amount of training data from HNZ is reduced. Just like Figure 1, the smaller the amount of training data is used, the bigger gain cTag features provide (e.g., over 4% absolute tagging accuracy gain over the baseline when 10% of training data is used for training).

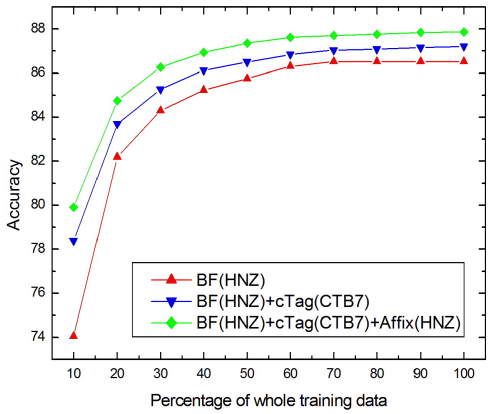

Figure 2: Performance of the POS tagger with three feature sets. Basic and affixes features are extracted from a reduced amount of training data in HNZ; cTag features are extracted from CTB7.

[13] One minor difference from word segmentation experiments is that using cTag features collected from both corpora slightly outperforms the system where cTag features are collected from CTB7 only (#7 vs. #5; #11 vs. #9).

Expt id	Basic Features (BF)		Additional Features			Tagging accuracy
	BF(HNZ)	BF(CTB)	cTag(HNZ)	cTag(CTB)	Affix(HNZ)	
#1	+					86.44
#2		+				77.03
#3	+	+				85.60
#4	+		+			86.86
#5	+			+		86.96
#6	+		+(u)	+(u)		86.93
#7	+		+(f)	+(f)		87.10
#8	+		+		+	87.27
#9	+			+	+	87.53
#10	+		+(u)	+(u)	+	87.38
#11	+		+(f)	+(f)	+	**87.65**

Table 11: Performance of the POS tagger with different feature sets and training data. The meanings of "+", "(u)", "(f)" are the same as in Table 9. The last column shows tagging accuracy; Accuracy of the baseline system is underlined, and the best result is in boldface.

5.4. Discussion

A few points are worth emphasizing. First, labeled data in Modern Chinese can help improve performance of NLP systems for Archaic Chinese, but only when it is used wisely. Using the most frequent cTags of characters based on frequencies collected from CTB7 gives a boost, whereas simply adding CTB7 to the training set hurts system performance. To understand this result, we look at characters that appear in both corpora. There are 2,704 such characters, accounting for 60.1% of character types in CTB7 and 68.3% in HNZ. From each corpus, we compile a list of (char, freq-ctag) pairs, where *char* is a character that appears in both corpora, and *freq-ctag* is the most frequent cTag of *char* in that corpus. It turns out that 54.0% of pairs in the two lists are identical. For comparison, we also compile a list of (char, freq-position-tag) pairs from each corpus, where *freq-position-tag* is the most frequent position tag (i.e., *S, B1, B2, B3, M, E* as used in word segmentation) of a character in a corpus. The agreement of these two lists is only 26.0%. That explains why cTag features from CTB7 help, while basic features from CTB7 hurt, the performance of the word segmenter.

Second, the simple heuristics used to infer the cTag of a character in a word are error-prone, and the cTag gathered from a small amount of data could be unreliable. Adding cTag features under that setting could hurt performance (e.g., Expt #4 in Table 9). Once the training set is reasonably large (even when the data is from a very different corpus such as CTB7), adding cTags always provides a boost. In contrast, due to the nature of DLG features (i.e., preferring to treat frequent strings in a corpus as a word), collecting DLG scores from HNZ works better than from CTB7. Therefore, when multiple corpora are available, which corpus should be used for a particular feature depends on the nature of the feature.

Third, owing to the nature of cTags, which are inferred from the POS tags of the words, POS tagging gets a bigger boost from cTag features than word segmentation, especially when the amount of training data is small.

6. Conclusion

In this paper, we introduced a segmented and POS tagged corpus for Archaic Chinese, compared it with CTB7 in several aspects that are related to word segmentation and POS tagging. More importantly, we demonstrated that, despite the differences between the two corpora, adding cTag features from CTB7 can indeed improve performance of both word segmentation and POS tagging on Archaic Chinese, especially when the amount of training data from Archaic Chinese is small.

For future work, we plan to identify more features like cTags which stay relatively stable as the language evolves, and determine whether adding those features could improve system performance the same way that cTag features do. We would also like to explore whether such features can be identified automatically without prior knowledge of the language. Finally, we plan to apply our methodology to other languages (e.g., Middle and Modern English) and other tasks (e.g., parsing).

7. References

Bollmann, M. (2013). POS Tagging for Historical Texts with Sparse Training Data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 11–18, Sofia, Bulgaria, August.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)*, pages 256–263.

Erjavec, T. (2012). The goo300k Corpus of Historical Slovene. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2257–2260, Istanbul, Turkey, May.

Hendrickx, I. and Marquilhas, R. (2011). From Old Texts to Modern Spellings: An Experiment in Automatic Normalisation. *JLCL*, 26(2):65–76.

Kit, C. and Wilks, Y. (1999). Unsupervised Learning of Word Boundary with Description Length Gain. In *Proceedings of CoNLL-99*, pages 1–6.

Kit, C. (2000). *Unsupervised Lexical Learning as Inductive Inference*. Ph.D. thesis, University of Sheffield.

Kit, C. (2005). Unsupervised Lexical Learning as Inductive Inference via Compression. In Minett, J. W. and Wang, W. S., editors, *Language Acquisition, Change and Emergence*, pages 251–296.

Kroch, A. and Taylor, A. (2000). Penn-Helsinki Parsed Corpus of Middle English, second edition.

Kroch, A., Santorini, B., and Diertani, A. (2004). Penn-Helsinki Parsed Corpus of Early Modern English.

Liu, Z. (2010). Why Yi Construction is not Disposal Construction in Ancient Chinese. *Journal of Chinese Language History*, 10:132–143.

Marcus, M., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mei, T.-l. (1990). The Origination of Disposal Construction in Tang and Song Dynasties (in Chinese). *Zhongguo Yuwen*, 8(3):191–206.

Ogiso, T., Komachi, M., Den, Y., and Matsumoto, Y. (2012). UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 911–915, Istanbul, Turkey, May.

Rayson, P., Archer, D., Baron, A., Culpeper, J., and Smith, N. (2007). Tagging the Bard: Evaluating the Accuracy of a Modern POS Tagger on Early Modern English Corpora Abstract.

Rögnvaldsson, E., Ingason, A. K., Sigurdsson, E. F., and Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1977–1984, Istanbul, Turkey, May.

Scheible, S., Whitt, R. J., Durrell, M., and Bennett, P. (2012). GATEtoGerManC: A GATE-based Annotation Pipeline for Historical German. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3611–3617, Istanbul, Turkey, May.

Song, Y. and Xia, F. (2012). Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3853–3860, Istanbul, Turkey, May.

Song, Y. and Xia, F. (2013). A Common Case of Jekyll and Hyde: The Synergistic Effect of Using Divided Source Training Data for Feature Augmentation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP-2013)*, pages 623–631, Nagoya, Japan.

Song, Y., Klassen, P., Xia, F., and Kit, C. (2012). Entropy-based Training Data Selection for Domain Adaptation. In *Proceedings of COLING-2012*, pages 1191–1200, Mumbai, India, December.

Wang, L. (1980). *Hanyu Shi Gao [An Outline of Chinese Language History]*. Zhonghua Book Company, Beijing.

Xia, F. (2000a). The Part-of-Speech Guidelines for the Penn Chinese Treebank (3.0), IRCS Report 00-07. Technical report, University of Pennsylvania, Oct.

Xia, F. (2000b). The Segmentation Guidelines for the Penn Chinese Treebank (3.0), IRCS Report 00-06. Technical report, University of Pennsylvania, Oct.

Xue, N. and Shen, L. (2003). Chinese Word Segmentation as LMR Tagging. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 176–179, Sapporo, Japan, July. Association for Computational Linguistics.