International Journal of Knowledge and Language Processing
Volume 3, Number 1, **2012**

www.ijklp.org
KLP International ©2012 ISSN 2191-2734
pp. **54-63**

State of Art of Pre-Qin Chinese Information Processing —Case Studies with Mencius and its Annotations and Commentaries

Shehui Liang
International College for Chinese Studies
Nanjing Normal University
No.122, Ninghai Road
Nanjing, 210097, China
liangshehui@163.com

Received March 2012; revised April 2012

ABSTRACT. *The information age calls for automatic processing and application of Pre-Qin Chinese corpus based on resources of Pre-Qin documents. Using Mencius and its annotations and commentaries as case studies, this article summarizes the state of art of information processing with Pre-Qin documents, covering topics ranging from traditional researches, to sentence alignment, automatic word segmentation, part-of-speech tagging, and word sense disambiguation etc. In addition, analysis shows that the annotations and commentaries for Pre-Qin documents provide novel and feasible resources for information processing with Pre-Qin documents.*
Keywords: Pre-Qin Chinese; annotations and commentaries; information processing

1. **Introduction.** In the Pre-Qin period of Chinese history, our ancestors have created a splendid heritage of Chinese civilization. In this period, Confucius, Mencius and other Hundred Schools of Thought initiated the first cultural and academic prosperity in Chinese history. Many schools of thought, such as Confucian, Taoism, legalism, and Mohism came into being, together with master-pieces such as LunYu, Mencius, and ZuoZhuan etc. Successive dynasties have made enumerate researches on them, which is often of vital importance for the passing on of Chinese civilization.

With time passing by, these scriptures or biographies distinguished themselves from that of contemporary readers, and the language in Pre-Qin documents became unreadable to their offspring. Therefore, the books of annotation books appeared, devoted specially for the explanation of these Pre-Qin documents. Still later, Chinese language changed with age and people could not even understand the "annotations", which made it necessary to annotate, elaborate and convey the messages encoded in the old annotations. This led to "comments", the explanation and elaboration of the old annotations. It can be said that over the thousands of years, China has made great achievements in the researches of scriptures

and their related annotations. These annotations and commentaries not only protected the scriptures from being lost, but also promoted the development of the studies of the classics and the progress of civilization.

The development of modern computer hardware and software, together with the progress made in Chinese information processing, make the information processing on Pre-Qin documents (including Mencius) and their annotations and commentaries possible. Tentative exploration has been made on sentence alignment, automatic word segmentation, speech tagging, and word sense disambiguation, which are different from modern Chinese information processing methods and means.

2. **Traditional researches on Mencius and its Annotations and commentaries.** Since ancient times, the experts and scholars have made many achievements in researches of the Pre-Qin literatures including Mencius. In[1] , there were not annotations for Mencius in the Western Han Dynasty (206 B.C.-A.D. 24). But in the Eastern Han Dynasty, people began to pay more attention to Mencius and show higher regard towards him. Thus monographs on Mencius appeared. It was said in legend that there were five different versions of annotation on Mencius — two versions of *Mencius Chapters* by Cheng Zeng and by Gao You, and two versions of *Notes on Mencius* by Zhao Qi and by Liu Xi. Until now, all the monographs are lost except the *Mencius Chapters* by Zhao Qi, which are the only intact and earliest annotations to Mencius in Han dynasty. Therefore, it is the only reliable reference for Mencius studies in Han Dynasty and of high historical value.

Throughout the period of Three Kingdoms, there were no new researches on Mencius. And the only one monograph in the Jin Dynasty is the *Mencius Notes* by QI Wusui. However, there was a large scale of development in the studying and interpreting Mencius during the Song dynasty. There were more than 100 types of studies and over 20 types are preserved and passed on to the present. *Annotations and Commentaries on Mencius*, by Sun Shi, is the first recorded annotation on Mencius in the North-Song Dynasty and also the earliest Mencius annotations circulated among the scholars and officials. Zhu Xi is the master of Neo-Confucianism. The book of Comments and Annotations on Shishu contains the essence of Zhu Xi’s neo-Confucianism. The section on Mencius represents the highest academic achievements of Zhu Xi’s research on Mencius, and the highest level of study of Mencius in neo-Confucianism in the Song Dynasty.

There were large number of monographs on Mencius in Yuan and Ming Dynasty, most of which were discussions on theories and ideas and senses of words focused on the book of The Mencius Annotations by Zhu Xi, with few innovations and developments.

In Qing Dynasty, the trend towards academic pragmatics and political security under high pressure forced the Qing academics to tread on the way of finding proofs for classics and thus characterized with textual exegesis on ancient academics. During the Qing dynasty, great importance was attached to the research on annotations of Mencius of the Han. There are many famous works on the annotations by Zhao QI, such as *Annotation on Mencius* by Jiao Xun, *Mencius Zhao Notes Correction* by Song Xiangfeng, and *Mencius Zhao Notes Research* by Gui Wencan etc. Among works, the most influential is the Annotation on

Mencius by Jiao Xin, which is his masterpiece. This book has been very influential in the studies of Mencius in Qing Dynasty.

The Mencius Interpretations by a modern person Yang Bojun[2], made meticulous annotations and translations on Mencius. Each chapter was divided into three parts: the original text, the annotations, and the translation. What should be mentioned is that in the book there was a "Mencius Dictionary", which is a very useful facility for research on Mencius and even information processing.

3. Research on sentence alignment technology. Currently, alignment is generally regarded as one of the indispensable steps in bilingual parallel corpus construction, and also one of the key steps in machine translation and bilingual dictionary compiling. The alignment of bilingual corpus can be divided into the following phases: passage alignment, sentence alignment, paraphrase alignment, and word alignment etc. The sentence alignment of parallel corpus is a process of establishing a correspondence between a set of sentences in the source language and the ones in the target language according to the contents of sentence content. Currently, the methods used in sentence alignment mainly include length-based sentence alignment method, dictionary-based sentence alignment method, length and dictionary-based sentence alignment method.

The features of length-based sentence alignment methods are: treat the sentence alignment as functions of sentence length; need no additional dictionary information. The method is weakened by the spread of faults. In China, the researches on alignment algorithm is firstly found in [3], who makes use of length-based method for a initial alignment of texts, and then identifies the anchor point in the bilingual parallel text and automatically extract the key words corresponding to the bilinguals so as to lower the complexity of alignment and reduces the spread of mistakes. Finally, we get a sentence alignment by utilizing the vocabulary alignment information.[4] realizes the alignment between English and traditional Chinese by making use of length-based method of Gale and Church. Moreover, he aligns sentences by combing the length method and vocabulary method through the using of special word list with date and mechanism. However, this method is not portable. Finally, [5] put forward the translation-based automatic alignment of bilingual sentences, the basic idea of which is adopting the bilingual dictionary as a bridge: to find the corresponding translations in the dictionary according to the words in the English sentences, and then match the translations to the Chinese sentences; to find the alignment sentence pair according to the evaluation function and dynamic programming algorithm.

The dictionary-based alignment algorithm also adopts the dynamic programming algorithm, and the only difference is that the finding of maximum probability is replaced by the finding of the evaluation function of each sentence pair.

Most of the early sentence alignments adopted the length-based method, which hypothesizes that there is a direct proportion between the source language and the target language. This is also the method adopted by Gale and Church [6], in which they defined the sentence length as the characters in the sentences, and evaluated the alignment program

between two sets of sentences by using the sentence length. This kind of algorithm aims at building an alignment relationship between sentences with approximate length. [7] adopt an offset alignment-based method: firstly you need a small-scaled dictionary, which provides some aligned base points. Each word is correspondent to a signal, with position vector signifying its position in the text and reach vector signifying the word counts between different positions of the word. If there is both a small difference of frequency of occurrences and positions among the words of two different languages, then we can use the dynamic programming algorithm to calculate their similarities. We choose the words of great similarity to form the bilingual dictionary, and then mark out the aligned word pairs and finally, search for the alignment between the source language and the target language by making use of the dynamic programming method again.

[8] considers the sentence length, Chinese characters and punctuations comprehensively, and put forward the model of translation between the ancient and modern Chinese sentences, realizing the automatic sentence alignment of ancient and modern Chinese based on genetic algorithm and dynamic programming algorithm. This is one of the few articles dealing with sentence alignment related to the ancient Chinese. [9] put forward the auto discovery method of the ancient alternative versions: firstly the similarity of sentence beads are obtained, and then the most possible sentence bead pair is found according to the similarity, then by constantly deleting the longest "identical text" in the sentence bead of the alternative versions, the alternative texts are identified. The case study is *Three Commentaries on the Spring and Autumn Annals*, and the results indicates that the sentence bead pairs are all correct and the matching algorithms can find correctly the alternative texts all conformed to the definitions.

4. Researches on Automatic Word Segmentation. Word segmentation is a process of recombining serial character sequence into word sequence. In English, the words are separated by the blank space as the natural delimiter, but in Chinese, only the characters, sentences and passages can be separated simply through obvious delimiters except the words that do not have a formal delimiter. Although there are similar questions in dealing with phrases in English, Chinese segmentation is more complicated and difficult than English. Therefore, the automatic word segmentation is one of the basic technologies and also difficult points of Chinese information processing.

Current word segmentations dealing with modern Chinese text can be summarized into three classes: string matching-based word segmentation, understanding-based word segmentation, and statistics-based word segmentation. String matching-based word segmentation is also called mechanic word segmentation, which refers to matching the character strings to be analyzed to the entries of a "sufficiently large" machine dictionary according to some strategies. If you find a certain character string in the dictionary, then the successful match shall result in the recognition of a word. According to the differences of scanning directions, string match word segmentation can be divided into positive match and reverse match. According to different length priority, they can be divided into maximum match and minimum match. As to dictionary-based word segmentation, there are several

factors, which influence its accuracy: (1) the choice of headword and the number of entries in the machine dictionary; (2) segmentation ambiguity; (3) unlisted word; (4) word segmentation method. The influence of the dictionary on the segmentation accuracy is even far greater than the ambiguous segmentation errors and the unlisted words produced by segmentation method itself.

The understanding-based segmentation method is also called artificial Intelligence, in which the recognition of words is achieved through the computer's imitating human's understanding of sentences. In recent years, some hot issues in artificial intelligence are applied to segmentation methods, and thus creating the expert system segmentation and neural network segmentation, both of which have their own weaknesses. For example, the drawbacks of the expert system are as follows: the incapability of learning from experience, the difficulty in maintaining a large knowledge base, and wasting too much time in segmentation of ambiguous multi-fields.

The statistics-based segmentation method is also called no dictionary segmentation method. The basic ideas are: words are the combination of steady characters, therefore in the context, the more co-occurrences of the adjacent characters, the more possible to form a word. So the frequency or probability of the adjacent co-occurrence among words can preferably reflect a word's reliability. We can count the frequency of the adjacent concurrent combinations of words in the corpus and calculate their co-occurrence information, which indicates the closeness of the combinations among the Chinese characters. When the closeness extent exceeds a certain threshold, and then comes to the possibility of the combination of characters forming into a word.

The advantages of the statistics-based segmentation method lie in its effectively automatic function of excluding ambiguity, the recognition of the new words and strange words, i.e. names and places etc. and its solving the drawbacks of dictionary-based segmentation method. But this method also has some limitations, for example, the frequent extraction of some high co-concurrent affirming groups, which are not words; its poor identification accuracy of common words and also it spends a lot of time and space.

Currently, there are probably only a few articles of word segmentation about ancient books, and research on Mencius has also not been widely published.[10] have made a research on 21 kinds of typical corpus of ancient Chinese including Mandarin and The Book of Lord Shang. By adopting the Great Chinese Dictionary as the automatic segmentation dictionary, they put forward a vocabulary treatment and quantification statistics based on computer automatic segmentation, and also a specific way of realization. However, they just vaguely come to a conclusion—there has been a steady increase in the proportion of ancient Chinese disyllables since Pre-Qin, which is accordant to the conclusion of the research on ancient Chinese vocabulary. In particular, [11] by adopting Conditional Random Fields, have made a comparative experiment—automatic segmentation and speech tagging on ZuoZhuan, which comes to the conclusion that the model "2w+2+C1", which is based on two words in context, bigram and character classification bigram, best suits the segmentation of ZuoZhuan.

5. POS Tagging. POS tagging is to mark out the category of each word of the text in specific context (nouns, verbs, and adjectives etc.), so it is also called category tagging. The importance of speech tagging lies in the fact that it is able to lay a foundation for syntactic analysis by confirming the grammatical function, facilitate the retrieval of syntactic structure in the speech tagging corpus and provide support for the homophone tagging, polyphone tagging and semantics tagging etc.

There are mainly three categories of the speech tagging methods dealing with modern Chinese: rules-based method; statistics-based method; rules and statistics-based method. The rules-based automatic speech tagging method first appeared in the 1960s. Along with the establishment of corpus, some scholars tried to make a machine automatic speech tagging on the English corpus, and gradually established a series of rules-based methods, among which the most representative was the tagging system developed in 1971. This method didn't possess strong robustness in natural language processing, and its accuracy rate couldn't meet the practical requirements.

In 1980s, under the influence of empiricism, the statistics-based method was gradually applied to corpus speech tagging, and then became dominant. The basic ideas of the statistics-based methods were: formulating part-of-speech mark set, selecting partial natural corpus to have an artificial speech tagging, obtaining statistic rules through calculating by using statistic theory, and then building a statistic model in the light of statistic rule, according to which the machine conducting a speech tagging. The biggest difference between the rules-based method and the statistics-based method lies in that in the former method, what the computer depends on is the artificial made linguistic rules, while in the latter, the computer automatically produces rules by depending on plentiful natural corpus.

There were both advantages and disadvantages of the rules-based method and statistics-based method, which couldn't meet the practical needs very well. So people began to adopt a combined method, i.e. combining the two methods to make up for each other's deficiencies, and adopting certain linguistic rules while building the statistic model by making use of large scale corpus. The experiment showed that this kind of compromised method had indeed improved the correct rate and work efficiency of the machine automatically speech tagging.

As for POS tagging on Ancient Chinese, the only research can be found is [11], which adopts CRFs for an integrated processing of segmentation and POS tagging after initial analysis and investigation on the document. Currently we haven't seen any other articles studying the speech tagging of the ancient books except the comparative experiment by Shi Min and Li Bin mentioned above. However, regretfully, the method used by Shi Min is still the speech tagging which is used to deal with the modern Chinese text.

6. Research on word sense disambiguation. Word sense disambiguation, the major work of which is to discriminate the different senses of the same word, aims at confirming the sense of the polysemic words in specific context.

[12] has comprehensively summarized the research achievements on word sense disambiguation in recent the 40 years, and made a discussion and comparison among

different disambiguation methods, including the method of selecting the most common meanings, using part of speech to make a sense disambiguation, the selection restriction-based method, word sense disambiguation of robustness, guidance way of learning, self-reliable sense disambiguation, no guidance way of learning and dictionary-based sense disambiguation etc.

[13] puts forward the context calculation model RFR_SUM (sum of relative frequency ratio) based on words collocation strength calculation, which is used to deal with the disambiguation of kinds of word levels. The RFR_SUM model has got a satisfactory result in sense disambiguation and the disambiguation of multi category words in Chinese information processing. The research, focusing on the disambiguation of kinds of word levels, has made outstanding achievements in the sense disambiguation of natural language processing as a basic issue.

In spite of the great achievements in natural language processing mentioned above, it seems that it's still difficult to judge the meaning of "pen" as "game fence" in "the box was in the pen", put forward by Bar-Hilell in 1956, with various kinds of methods of scholars. Thus it can be seen that it's still difficult in dealing with the word sense disambiguation.

[14] argues that we need to make more efforts in the following two aspects: 1) to continue to build a sense tagging corpus of large scale and high quality, which is very important especially for Chinese, because the tagging scale of English is 200,000, while that of Chinese is less than 100,000(in the fine grit). Without a large –scaled sense tagging corpus, it will be impossible to have a further development of many researches such as model training and evaluation contest, not to mention to obtain a sense disambiguation system. 2) To bring in more knowledge to improve the property of sense disambiguation. Such knowledge includes linguistic knowledge such as grammatical structure and sense similarity etc., and also large amounts of untagging corpuses, especially the Internet corpuses, such as Wikipedia. The latter needs a semi-guidance and weak-guidance model. From current research status, [14]'s opinions are surely reasonable.

There is a total difference between the sense disambiguation researches on ancient Chinese and modern Chinese. Currently, [15] is only one paper on ancient Chinese sense disambiguation. In this article, she firstly analyzes the distribution and its features of the ancient Chinese sememes, investigates the difficult points of sense disambiguation. Then based on current theories and methods of disambiguation and by using the conditional random field, which is based on statistic model of machine automatic learning, she selects the complex feature of the word and its part of speech in the context and puts in other appropriate linguistic features, designing six different templates, and then making a sense disambiguation experiment of the high frequency words in ancient Chinese such as "将", "如", "我", "信", "闻", and "之". The experiment has adopted the word sense disambiguation method, which is the same as the method of dealing with modern Chinese texts.

7. Conclusions. In conclusion, scholars of archeography, Exegesis and Philology have accumulated fruitful researches on the Pre-Qin documents including Mencius and their

annotations. The researchers of modern Chinese have also accumulated a wealth of experience on automatic word segmentation, speech tagging and word sense disambiguation. However, regretfully, there is no research on Mencius based on Chinese information processing both at home and abroad. There are also no researches on Mencius from the aspects of sentence alignment and annotation alignment between the original and citations, nor researches on automatic word segmentation as well as word sense disambiguation and its methods. Moreover, research on Mencius annotations and literatures in relation to information processing is an unexplored field.

Nowadays, the various annotations and literatures on Mencius in history mentioned above provide a totally new way and great resources for us to explore the information processing of Mencius. We have made a preliminary research on the features of Mencius and its annotations and literatures, and have obtained a better result in making an exploration of the features, difficulties, models and strategies of the information processing of the Pre-Qin literatures including Mencius. For example, we have made an automatic segmentation experiment on all of the characters of the 7th chapter of Mencius LiangHuiWang Chapters by using Mencius Zhushu, and the statistic results are as follows:

FIGURE 1: Experiment Result by Using Annotation Segmentation

Number	Statistical Item	Figure and Percentage
1	term frequency of artificial proofreading	2287
2	term frequency of machine tagging	2200
3	term frequency of machine tagging errors	252
4	accuracy rate of machine tagging	88.5%
5	recall rate of machine tagging	85.2%

The annotations can provide a knowledge source, which is far more reliable than the analysis obtained from the statistical models, for the automatic word segmentation and sense disambiguation of the pre-Qin literatures including Mencius. The preliminary test results show that we've got a correct idea, advanced technology and feasible method. The details will be dealt with in another article.

Of course, we still need to make some efforts to solve the problems in information processing of the pre-Qin literatures like Mencius. For example:

1. While studying the sentence alignment and annotation alignment of Mencius and its annotations and literatures, we'll come to complicated problems such as the discordance of the sentence fragments between the original text and the citations, off word and redundancy, which will interfere with the correct accuracy of information processing.

2. We have to make a collection, organization and proofreading of the electronic versions of Mencius and its annotations, and develop appropriate word segmentation rules and Part-of-speech tags set, as well as conducting automatic word segmentation, speech tagging, and word sense disambiguation of Mencius, all of which deal with confused content and huge workload.

3. What we use in processing Mencius and its annotations are Complex formed Characters, and even the characters that we cannot find in the Computer commonly used word stock.

4. Currently, there are only a few researches on the information processing of Mencius and other pre-Qin literatures, which results in a shortage of reference to theories and methods in this article. Moreover, due to their own features, the current information-processing model of modern Chinese is not suitable for the processing of Mencius and the related literatures.

In the information age of the 21st century, what we are concern with is trying to find a new method and a new idea of information processing in the automatic word segmentation and sense disambiguation of the annotation-based Mencius, which is based on the method of information processing and computational linguistics by means of the hi-tech computer. Moreover, we can conduct an experiment of automatic word segmentation and sense disambiguation at Mencius, through which making a deeper manufacturing and processing to meet the needs of the rapid progress of informational era. Exploring the application of modern computer processing technology to the automatic word segmentation and word sense disambiguation of Mencius and the whole pre-Qin literatures, and conducting an automatic word segmentation and word sense disambiguation on Mencius, etc., all of the above is a blank space that is waiting for us to fill in.

8. Acknowledgment. This Work is supported by Jiangsu Provincial Department of Education Foundation of Philosophy and Social Science (Grant No.2011SJB740010), Natural Science Foundation of the Jiangsu Higher Education Institution of China (GrantNo.11KJD520009), under the Priority Academic Program Development of Jiangsu Higher Education Institutions（PAPD）.

EFERENCES

[1] Dong Hongli. *Mencius Research*. Jiangsu Ancient Book Publishing House, NanJing, 1997.

[2] Yang Bojun. *Mencius Notes*. Zhonghua Book Company, Beijing, 1962.

[3] Liu Xin, Zhou Ming, Huang Changning. An Experiment of Chinese-English bilingual text alignment based on length algorithm // *Computational linguistics and applied research progress*. Tsinghua University Press, Beijing, pp.62-68, 1995.

[4] Wu, Dekai.Aligning a Parallel English Chinese Corpus Statistically with Lexical Criteria// 32nd Annual Meeting of the Association for Computational Linguistics, New Mexical, USA, pp.80-87, 1994.

[5] Qian Liping, Zhao Tiejun. Translation－Based Automatic Alignment of English and Chinese Parallel Corpora. *Computer Engineering and Applications*, Vol.36, No.12, pp.59-61, 2000.

[6] Gale, W. A. and Church, K. W. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, Vol.19, No.1, pp.75-102, 1993.

[7] P.Fung and K.McKeown.Aligning noisy parallel corpora across language groups: Word pair reature matching by dynamic time warping//AMTA-94, *Association for Machine Translation in the Americas*,

pp.81-88, 1994.

[8] Guo Rui, Song Jihua etc. Ancient Sentence Search Based on Sentence Auto Alignment in Parallel Corpus of Ancient and Modern Chinese. *Journal of Chinese Information Processing*, Vol.22, No.2, pp.87-91, 2008.

[9] Xiao Lei, Chen Xiaohe. Automatic Detection of Version Differences among Ancient Chinese Texts. *Journal of Chinese Information Processing*, Vol.24, No.5, pp.50-55, 2010.

[10] Qiu Bing, Huang Fujuan. Study on the Trend of Ancient Chinese Words Based on the Word Automatic Segmentation. *Microcomputer Information*, Vol.24, No.24, pp.100-102, 2008.

[11] Shi Min, Li Bin, Chen Xiaohe. CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese. *Journal of Chinese Information Processing*, Vol.24, No.2, pp.39-45, 2010.

[12] Feng Zhiwei. The Approaches for Word Sense Disambiguation (WSD). *Terminology standardization and information technology*, Vol.9, No.1, pp.31-375, 2004.

[13] Qu Weiguang. *Research on the Automatic Resolution of Word Level Ambiguity on Modern Chinese*. Science Press, Beijing, 2008.

[14] Jin Peng. A Brief Introduction to Word Sense Disambiguation. *Terminology Standardization and Information Technology*, Vol.13, No.3, pp.29-34, 2010.

[15] Yu Lili. The Ancient Chinese Word Sense Disambiguation Based on CRF. *Microelectronics and Computer*, Vol.26, No.10, pp.45-48, 2009.