International Journal of Knowledge and Language Processing
Volume 5, Number 4, **2014**

www.ijklp.org
KLP International ⓒ2014 ISSN 2191-2734
pp.1—13

The Study of the Automatic Word Segmentation Method in the Pre-Qin Chinese Corpora

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

Received June 2014; revised October 2014

ABSTRACT. *Automatic word segmentation is a basic issue in Chinese information processing. This article mainly studies the Conditional Random Fields Model-based method of automatic word segmentation and the automatic word segmentation methods by using relevant annotations and commentaries on Mencius. The results of the experiments of automatic word segmentation manifest that these two methods have a remarkable effect as the F value of words and sentences has achieved a relatively high level. When using different methods for automatic word segmentation of Mencius, in addition to the F value of words, this article tries to introduce the F value of sentences as another statistical index.*

Keywords: Pre-Qin literature; Mencius; automatic word segmentation; The Conditional Random Field Model; annotations and commentaries

1. Introduction. As one of the basic issues in Chinese information processing, the automatic word segmentation has made great progress in the last 20 years, and several mature automatic word segmentation systems have been designed and utilized [1]. Actually, a comparatively mature automatic word segmentation system can both provide a foundation and data support for further language researches, and serve as a prerequisite for constructing the systems of syntactic analysis, machine translation, and information retrieval. In this sense, the automatic word segmentation is also important for the Pre-Qin literature's information processing.

In terms of the digitalization of ancient books, in order to build up an effective corpus of ancient books, it is necessary to process the corpus in depth, and the automatic word segmentation is a critical step for the corpus processing. The linguistic data of ancient books which have gone through the automatic word segmentation and syntax tagging, is a valuable resource for researchers to gain knowledge, not a mere piling up of the texts.

Therefore, the reusability and utility of the documents have been improved. On the other hand, for the Pre-Qin and other ancient Chinese literature, the highly accurate automatic word segmentation of the original literature is conducive to helping us elaborate our researches in language style and features.

At present, the automatic word segmentation of the Pre-Qin and other ancient Chinese documents has achieved some results. Qiu Bing and Huang Fujuan have proposed an illuminating method——mixing word segmentation——in the word segmentation research of 21 kinds of ancient Chinese literatures including GUOYU and *The Analects of Confucius* [2]; Shi Min and Li Bin et al. conducted a meticulous word segmentation experiment on The Commentary of Zuo by using Conditional Random Fields (CRFs) and adopting complex Chinese characters' features as the template characteristic, and the open test F value has reached 94.60% [3]. Xu Runhua and Chen Xiaohe have considered the abundant language resources existing in the annotations and commentaries， and they used a large amount of language information to conduct the automatic word segmentation experiment on the quotations, as well as achieving a comparatively good result [4].

The work mentioned above has provided a reference to the exploration of the automatic word segmentation methods. Therefore, on the basis of/according to the longitudinal comparison of the existed methods, we deliberated globally and tried to carry out the automatic word segmentation experiment on Mencius based on the statistical models and by utilizing the documents of annotations and commentaries. The experiment and the result can provide the other Pre-Qin documents' automatic word segmentation with some reference.

2. **The Automatic Word Segmentation Method Based on the Statistical Models.** The Automatic Word Segmentation Method based on the statistical model, also known as the Automatic Word Segmentation Method based on probability， is to establish a statistical model for Automatic Word Segmentation so as to obtain each parameter of the model through machine learning of the labeled language materials, and then pick up the string of words which is in the highest probability among the various possible word strings as the output result [5]. According to the recent years' SIGHAN competition reports, the automatic word segmentation method based on the statistical model engendered remarkable effects. [1] Firstly, this article tests whether the automatic word segmentation method based on the statistical model can work well on the *Mencius*, secondly, it tests the adaptability of the method after we added the features of ancient Chinese documents.

2.1. **The Conditional Random Fields Model.** At present, the Automatic Word Segmentation Methods based on the statistical models include the N-grammatical model, Hidden Markov Model, Maximum Entropy Model and the Conditional Random Fields Model. Among these, the Conditional Random Fields Model has its own advantages in terms of automatic word segmentation task: by learning language materials, the Conditional

[1] http://sighan.cs.uchicago.edu/bakeoff2006/longstats.html

Random Field does not normalize at every local node but treats all the features in a global fashion, having the competence in expressing the features of elements' long-distance dependency and overlapping, which compared with the N-grammatical model and the Hidden Markov Model, has more experimental value.

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum \lambda_k f_k(y_{i-1}, y_i, x) + \sum \mu_k g_k(y_i, x))$$

The so-called Conditional Random Field denotes such an undirected graphical model that under the condition of the given node (the observed value), it will calculate and output the node (the label)'s conditional probability. The formalized definition is:

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum \lambda_k f_k(y_{i-1}, y_i, x) + \sum \mu_k g_k(y_i, x))$$

Every $f_x(\ )$ is the feature of the output nodes in position i and i-1 in the observed sequence x, and every $g_x(\ )$ is the feature of both the input and output nodes in position i. Besides, λ and μ are the characteristic function's weights, and Z is the normalization factor.

The Conditional Random Fields Model used in our experiment has adopted the toolkit −"CRF++0.50" programmed by TakuKudo to train and test[2].

2.2. **The CRFs Word Segmentation Principle.** There have been several studies on the utilization of the Conditional Random Fields Model in the Automatic Word Segmentation experiments. The basic principle is changing the Automatic Word Segmentation problem into the problem of the lexeme information's sequence tagging. Usually, Chinese characters sequence can be labeled as four-word marks, six-word marks，etc. In four-word tags，for instance, the Chinese characters sequence {孟 子 見 梁 惠 王} can be labeled as {B, E, S, B, M, E} correspondingly.

The Conditional Random Fields Model-based word segmentation system can be divided into two parts—the training part and the testing part. The former uses the feature template to extract the features from the corpus, and then interactively learns the weights by the algorithm. The latter identifies the character sequence tags by making use of the features and their parameters from the training part, and finally, reconstructs the words in accordance with the word tags.

Apart from that, the *Mencius* and other Pre-Qin documents have unique language characteristics and character features. We may consider to add the already-known ancient Chinese characters' features into the statistical model and conduct experiments to see the effect.

To sum up, we adopted the Conditional Random Fields Model to do the full-text automatic word segmentation experiment on the *Mencius*. With consideration of the basic literal information and the features in the usage of the ancient Chinese characters, we have taken the other linguistic features such as the initial (consonants), final (also rhyme, mainly composed by vowels or the combinations of vowel and nasal consonant), tone, radical etc.

[2] http://crfpp.sourceforge.net/

into account. Moreover, we planned to introduce *The Commentary of Zuo* and *The Analects of Confucius* as the training language materials into the experiment.

2.3. **The CPFs-Based Word Segmentation Experiment.** The training materials of our adopted experiment include the *Mencius*, *The Analects of Confucius* and *The Commentary of Zuo* which are in similar length with about 22,000 characters for each word segmentation text. The testing materials adopted include the Mencius, whose text length is about 9.500 characters. There is no text cross between training and testing materials, both of which are open tests, and the corpus scale ratio is 7 to 3.

The feature selection is of great significance to the machine learning and the statistical model-based automatic word segmentation. It influences not only the language model's speed performance, but also the model's accuracy and adaptability. Thus, the CRFs-based automatic word segmentation experiment entails careful consideration of how to select the proper features in the light of the constitution of the ancient Chinese language materials. We chose the simple literal information and the complex Chinese characters features as the text features for the automatic word segmentation experiment so as to select the best feature.

The simple literal feature is determined by the different window lengths of the character string's observed sequence and the character's co-occurrence. In this article, "W" stands for character, "nW" indicates the window [-n, n]'s character, and "2W" means the contextual window [-2,2]'s character. Besides, "nW+m" signifies the characters' co-occurrence, and "+2" denotes that two characters co-occurred.

The complex Chinese characters features are decided by the Chinese characters own information or the information of the manual tagging. The complex Chinese characters features adopted by this experiment include: (1) the character classification: we categorized them into the Chinese character (HZ), common punctuation (Punc), punctuation at the end of sentence (SenPunc) and the Chinese character number (CNum) in order to strengthen the pertinence of model learning, expecting to improve the classification accuracy; (2) the sound, rhyme, tone and radical information: since the initial, final, and tone in Pre-Qin are reconstructed and there are no publicly recognized data base, we selected the word table (or character table) of *GuangYun* which describes the middle Chinese as the approximate basic data base, and the radical information referred to the historical books and dictionaries' radical classification and was labeled by manual work.

Using the simple literal information as the feature, Table 1-4 compared the results of the contextual windows from left to right with 2-3 characters, and the two/three characters' co-occurrence situation. As the result shows, the *Mencius* as the training language material turned out to be the most effective. Thus, it can be seen that the machine learning can adapt to the ancient Chinese comparatively well. In addition, both *The Analects of Confucius* and *The Commentary of Zuo* gained a good result as well. The word F value surpassed the Baseline, but there existed a wide gap in sentence F value compared to the Mencius as the training language material. Nevertheless, the result is better than *The Commentary of Zuo*, which statistically confirmed that the Mencius and The Analects of Confucius are in high

homogeneity.

TABLE 1: THE *MENCIUS* TRAINING, CLOSE TESTING RESULT

Template	**2W**	**3W**	**2W+2**	**2W+3**	**3W+2**	**3W+3**
Word Recall Rate	0.980	0.982	0.989	0.983	0.985	0.993
Word Accuracy	0.980	0.982	0.989	0.983	0.985	0.993
Word F Value	0.980	0.982	0.989	0.983	0.985	0.993
Sentence Recall Rate	0.902	0.909	0.937	0.902	0.909	0.956
Sentence Accuracy	0.902	0.909	0.937	0.902	0.909	0.956
Sentence F Value	0.902	0.909	0.937	0.902	0.909	0.956

TABLE 2: THE *MENCIUS* TRAINING, OPEN TESTING RESULT

Template	**2W**	**3W**	**2W+2**	**2W+3**	**3W+2**	**3W+3**
Word Recall Rate	0.942	0.943	0.953	0.943	0.945	0.982
Word Accuracy	0.945	0.945	0.953	0.949	0.947	0.985
Word F Value	0.944	0.944	0.953	0.946	0.946	0.984
Sentence Recall Rate	0.800	0.803	0.823	0.823	0.860	0.911
Sentence Accuracy	0.800	0.803	0.823	0.823	0.860	0.911
Sentence F Value	0.800	0.803	0.823	0.823	0.860	0.911

TABLE 3: *THE COMMENTARY OF ZUO* TRAINING, THE *MENCIUS* TESTING RESULT

Template	**2W**	**3W**	**2W+2**	**2W+3**	**3W+2**	**3W+3**
Word Recall Rate	0.862	0.870	0.903	0.889	0.910	0.922
Word Accuracy	0.868	0.880	0.901	0.890	0.914	0.919
Word F Value	0.864	0.875	0.902	0.889	0.912	0.920
Sentence Recall Rate	0.580	0.582	0.600	0.602	0.611	0.609
Sentence Accuracy	0.580	0.582	0.600	0.602	0.611	0.609
Sentence F Value	0.580	0.582	0.600	0.602	0.611	0.609

TABLE 4: *THE ANALECTS OF CONFUCIUS* TRAINING, THE *MENCIUS* TESTING RESULT

Template	**2W**	**3W**	**2W+2**	**2W+3**	**3W+2**	**3W+3**
Word Recall Rate	0.872	0.940	0.923	0.912	0.910	0.931
Word Accuracy	0.878	0.950	0.921	0.920	0.914	0.931
Word F Value	0.874	0.945	0.922	0.916	0.912	0.931
Sentence Recall Rate	0.680	0.682	0.600	0.692	0.700	0.699
Sentence Accuracy	0.680	0.682	0.600	0.692	0.700	0.699
Sentence F Value	0.680	0.682	0.600	0.692	0.700	0.699

In order to acquire better word segmentation effect, we chose the “3W+3” which can segment words well in the experiment above as the base, added the complex Chinese character feature, and then conducted the CRFs word segmentation experiment on the *Mencius*. The concrete training language material sample is shown in Table 5.

TABLE 5: THE CRFS WORD SEGMENTATION TRAINING LANGUAGE MATERIAL SAMPLE

Character	col1 Character Classification	col2 Initial consonant	col3 Final (Rhyme)	col4 Tone	col5 Radical	Answer
孟	HZ	明	映	去	子	B
子	HZ	精	止	上	子	E
見	HZ	见	霰	去	見	S
梁	HZ	來	陽	平	木	B
惠	HZ	匣	霽	去	心	M
王	HZ	云	陽	平	玉	E
。	SenPunc	*	*	*	*	S

According to the frequency of use of the complex Chinese character feature, we did five groups of experiments respectively (as shown in Table 6). In the table, CN stands for adopting the nth list's complex Chinese character feature for the experiment.

TABLE 6: FIVE GROUPS OF THE RESULTS IN THE EXPERIMENTS

	3W+3+ C1	3W+3+ C123	3W+3+ C1234	3W+2+ C12345	3W+2+ C15
Word Recall Rate	0.980	0.979	0.980	0.976	0.980
Word Accuracy	0.986	0.985	0.982	0.980	0.984
Word F Value	0.983	0.983	0.981	0.978	0.982
Sentence Recall Rate	0.941	0.900	0.922	0.938	0.941
Sentence Accuracy	0.941	0.900	0.922	0.938	0.941
Sentence F Value	0.941	0.900	0.922	0.938	0.941

It can be found by analyzing the experiments' results that adding the character classification feature is conducive to enhancing word segmentation accuracy. Therefore, the result of template "3W+3+C1" is slightly higher than the first experiment, and the sentence F value has reached 0.941. Apart from this, adding the initial, final, tone and radical information to the character did not make any significant difference because the three features — initial， final and tone — themselves also need to be disambiguated. Every character has its own initial, final and tone, and they are always different under the condition of different word property or semantic items, which needs to be carefully studied further.

Although these conclusions are limited in the language material scale, as for the statistical model-based automatic word segmentation on the *Mencius*, the conclusion can be drawn from the experiments that the template "3W+3+C1", based on the three-Chinese-character context, three-Chinese-character co-occurrence, and taking the

character classification into account, is the most adaptive one for the automatic word segmentation on the *Mencius*.

3. The Automatic Word Segmentation Method by Using Annotations and Commentaries. The rule-based and statistical model-based automatic word segmentation methods can be viewed as the direct transfer of the modern Chinese word segmentation to ancient Chinese, which have not taken a full consideration of the Pre-Qin documents, esp. the large amount of language information and the word features of the correlative annotations and commentaries. Meanwhile, although these automatic word segmentation methods can get good experimental results, they still lag behind in speed and convenience due to the processes of designing the training language material, selecting the model's features, machine learning and so on.

Hence, we chose the *Mencius* and its annotations and commentaries as the object of the study, and discussed a word segmentation method which is based on the Pre-Qin literature's characteristics with adequate consideration of the specialty of the ancient Chinese information processing. This method's experimental word F value can achieve 92.6%, thus possessing practical value. Different from the former methods, this kind of word segmentation method does not need to make the word list or the training language material in advance, so it has a high universality and can be transferred to other Pre-Qin documents' word segmentation tasks.

3.1. **An Overview of the Automatic Word Segmentation Method by Using Annotations and Commentaries.** On account of that the Pre-Qin documents are age-old and the language at that time is quite elusive, the descendants have to read the annotations and commentaries to deepen their understanding of the original literature. As for the computers, these annotations and commentaries serve as very good language knowledge base, containing lots of information about every language unit, and thus can help the computer with machine learning and data mining. Taking Annotation on Mencius as an example, the machine can learn a large amount of the information on the word level so as to be the best reference for the automatic word segmentation on Mencius which are shown in Table 7.

From Table 7 we can see that the annotations have given language points' explanations to and word analyses in the original text. In other words, in the annotations, these language units have been divided and interpreted in a proper way, which is beneficial to the automatic word segmentation and POS tagger on the Pre-Qin documents. Xu Runhua, et al. posited that the Pre-Qin document information processing can take the annotations as the priming knowledge for further literature information processing such as the automatic word segmentation, POS tagger, syntactic analysis etc. [4]. Since this type of knowledge is often composed for a certain document, they are more pertinent, and to some extent, more reliable than the statistical models.

TABLE 7: THE INSTANCE OF MACHINE LEARNING

NO.	Mencius (the original text)	Annotation on Mencius (the annotation)
1	**莊暴**|見|孟子	**莊暴**，齊臣也。
2	有|為|**神農**|之|言|者|**許**|**行**	**神農**，三皇之君，炎帝神農氏。**許**，姓；**行**，名也。
3	**蒙**|學|射|於|**羿**	**羿**，有窮後羿。逢**蒙**，羿之家眾也。
4	人|不|知|亦|**囂囂**.	**囂囂**，自得無欲之貌也。

On the other hand, the annotations and commentaries have gone through the development of the past dynasties, so there are different versions for us to refer to. For example, *The Analects of Confucius* has several annotation-commentary documents of different periods, like *Annotated Confucian Analects*, *Pen-Conversation on the Analects*, *Annotated glossary of the Analects*, and *The whole interpretation of Analects of Confucius*. *The Commentary of Zuo*'s annotations and commentaries have experienced the evolution from Confucian classics (*Spring and Autumn Annals*) to biography (*The Commentary of Gongyang*) and then to exegesis and commentaries (*Annotated Commentary of Gongyang*). The famous annotations and commentaries of the *Mencius* are *Annotated Mencius*， *The whole interpretation of Mencius* and *The proper annotations on Mencius* etc. Among these, *The whole interpretation of Mencius* annotates the *Mencius* directly, which belongs to the type of using the Chinese in Song Dynasty to interpret the Chinese in Pre-Qin period (Old Chinese); *Annotated Mencius* and *The proper annotations on Mencius* comment on *the Mencius Chapters*, while *the Mencius Chapters* annotates the *Mencius*, which belongs to the type of using the Chinese of Ming Dynasty to interpret the Middle and Old Chinese.

From this angle, we have considered to use several versions of annotations and commentaries to conduct the Pre-Qin literature information processing. Actually, the biggest function by doing so is to complement each other and to correct the errors. If we cannot find some words' or sentences' interpretation in a certain annotation-commentary document, we may find them in the others, esp. the versions belonged to different dynasties. If there are inconsistent interpretations in two annotation-commentary documents, or some mistakes, we may often resort to the third version for identification and selection.

Based on this consideration, we respectively utilized *The whole interpretation of Mencius*, *Annotated Mencius* and *The proper annotations on Mencius* to conduct the automatic word segmentation experiments, and on this basis, we integrated the three annotation-commentary documents' information to improve the word segmentation effect. Because of the similarity of the three versions, the system will not endure much pressure, and this method has comparatively high efficiency and feasibility.

3.2. **The Automatic Word Segmentation Method by Using Annotations and Commentaries and Its Experiments.** Before automatically word-segmenting the *Mencius*, we have tried the automatic sentence alignment experiment between the *Mencius* and the

three annotation-commentary documents. The result is that the maximum sentence alignment F value has reached 0.9895, and the maximum annotation alignment F value has achieved 0.900. Now we used the annotation alignment contents in the annotation-commentary documents to produce three different annotation word lists, among which the word list of *Annotated Mencius* includes 1828 words, the word list of *The whole interpretation of Mencius* records 1136 words and the word list of *The proper annotations on Mencius* includes 1952 words. We also counted, as to the three annotation word lists, the proportions of the unlisted word type in *Mencius*: as for *Annotated Mencius*, the proportion is about 40%; as for *The whole interpretation of Mencius*, the proportion is about 44.5%; as for *The proper annotations on Mencius*, the proportion is about 40.1%. Besides, we tried to make simple statistics on the two annotation word lists of *Annotated Mencius* and *The whole interpretation of Mencius*, and we found that there are 527 same words in the two lists, 528 words only in the list of *The whole interpretation of Mencius*, and 1136 words only in the list of *Annotated Mencius*. Thus it can be seen that different annotation-commentary documents' word lists have a strong complementarity. After the generation of the word lists, we might use these lists to do the automatic word segmentation experiments.

The experimental method is that we adopted the rule-based automatic word segmentation algorithm as the main frame: every time we extract the character string whose length is equal to the maximum word length from the text which is to be segmented, and circularly compare it with the token word in the annotation word list, diminishing it word by word until we successfully find it in the annotation word list, or the residue character would be outputted as a one-character word and be manipulated in the whole text interactively.

3.2.1. **The Generation of the Annotation Word List.** It can be indicated from the concrete experimental methods that the core problem of the annotation-commentary document-based automatic word segmentation is that we must construct the annotation word list well for the system to use. Our train of thought is that we can transfer the construction of the annotation word list into the search of all the character strings which are all possible to form words in the original text, and add character strings into the annotation word list as words. This work is related to the annotation alignment.

When searching in the original text, we actually applied two kinds of string matching methods, the broad one and the narrow one. The so-called broad matching denotes if the character string of the original text appears in the annotations, and it is not included in the other character strings of the original text, then the matching is successful. The so-called narrow matching means if the character string of the original text appears in the annotations with the separation marks around (punctuation: “ ”, ‘ ’, 【】etc.; the guide words: 曰, 爲, 稱 etc.), then the matching is of success.

On the other hand, the range of the annotation also has an effect upon the matching result. In other words, the character string can not only match the whole passage of the successfully aligned annotations and commentaries, but also do the searching work in all the annotations and commentaries. Therefore, according to whether making use of the

alignment information, the search scale can be divided into the alignment search and the global search. The former one is defined as that every character string in the original text will be searched in the annotations or commentaries which are successfully aligned with the current original text. The latter one denotes that every character string in the original text will be searched in all of the annotations and commentaries.

By considering the two factors—the String Matching Method and the Searching Range of the Annotations Synthetically, we can generate four annotation lists for every annotation-commentary document. We can use these annotation word lists to construct proper word segmentation models, which can be applied to the automatic word segmentation experiments on the *Mencius*.

3.2.2. **The Design of the Word Segmentation Algorithm Model.** Different annotation word lists' generation methods will form different searching methods for the word lists. Generally speaking, the success of searching the word lists can decide whether the current character string will be cut into words. So, the crux of the word segmentation algorithm is how to judge whether every character string's word list searching is successful in the Mencius. The whole word segmentation algorithm is constrained by the factors such as the String Matching Method and the annotation's searching scale. In accordance with these factors, we have constructed an annotation-based automatic word segmentation algorithm model on the *Mencius* below:

$$RES = S_i \times R_j$$

S and R are arrays whose lengths are 2, i and j are array subscripts. The value of the array's elements is 0 or 1. The Res' value is 1 when the word list searching is successful, and it is 0 when the searching is failed. Among these, S_1 stands for adopting the narrow matching, and S_2 denotes the adoption of the broad matching; R_1 represents the adoption of the alignment searching, and R_2 signifies adopting the global searching. According to different values of i and j, the algorithm model has four parameter combinations, and every group of parameters forms a sub-algorithm, as shown in Table 8.

TABLE 8: THE WORD SEGMENTATION ALGORITHM MODEL'S PARAMETERS

	Algorithm1	**Algorithm2**	**Algorithm3**	**Algorithm4**
Matching	Broad	Broad	Narrow	Narrow
Alignment	Yes	No	Yes	No

We chose 10% of *Mencius*' language material and *The proper annotations on Mencius* to optimize the sub-algorithms, and the results are shown in Table 9. The word segmentation results show that the Sub-algorithm 1 is the most efficient and basically has the practical value.

TABLE 9: THE RESULTS AFTER OPTIMIZING THE FOUR SUB-ALGORITHMS

	Sub-algorithm1	**Sub-algorithm2**	**Sub-algorithm3**	**Sub-algorithm4**
Word Accuracy	0.923	0.917	0.917	0.910
Word Recall Rate	0.921	0.909	0.917	0.900
Word F Value	0.922	0.913	0.917	0.905
Sentence Accuracy	0.703	0.682	0.709	0.620
Sentence Recall Rate	0.703	0.682	0.709	0.620
Sentence F Value	0.703	0. 682	0.709	0.620

There are two main problems of the dividing mistakes according to the analysis. Firstly, only depending on the word list searching cannot solve the combinatorial ambiguity, which is quite common in *Mencius*' language materials, for instance, "可以", "可得"and so on. Secondly, some words are contained in another longer word, for example, "惠王" often appears in "梁惠王" , therefore, when "惠王" appears independently, the segmentation will be failed (because "惠王" is not included in the word list).

3.2.3. **The Automatic Word Segmentation Experiment on the Whole Text of the *Mencius.*** On the basis of the optimization of the model's parameters, we utilized three annotation-commentary documents as the basic experimental language materials to conduct the automatic word segmentation experiments. Meanwhile, by using the existing experimental results, we composed a "combining annotation" word list for testing. The comparisons of the word segmentation results from the experiment are shown in Table 10, from which we may find that combining the annotations has the best result. The main reason is that other annotations and commentaries have complemented the other's unlisted interpretations, for instance, there is no word like "東山" in *The whole interpretation of Mencius*' annotation word list, but they are in the *Annotated Mencius*' and *The proper annotations on Mencius*' .From the table we can see that *The proper annotations on Mencius* is more efficient than *The whole interpretation of Mencius* and *Annotated Mencius*. This is because that *The proper annotations on Mencius* is the longest in its length and its word list is the biggest on scale, moreover, as mentioned above, compared with the word lists of *The whole interpretation of Mencius* and *Annotated Mencius*, as for the word list of *The proper annotations on Mencius*, the word types of the unlisted words in the *Mencius* is less in amount. Thus, this would make, to some extent, the word list of *The proper annotations on Mencius* behave better. In addition, though there is no big difference between *The proper annotations on Mencius* and the combining annotations and commentaries, we still insisted on composing the annotations and commentaries in the main consideration of that *The proper annotations on Mencius* cannot record all the words, and the words' reliability in one single annotation-commentary document also remains to be proved by other annotations and commentaries, therefore, the work of combining the annotations and commentaries is also beneficial and more conducive to leak filling.

TABLE 10: THE ANNOTATION-COMMENTARY DOCUMENT-BASED AUTOMATIC WORD SEGMENTATION EXPERIMENTAL RESULTS

The Name of the annotations and commentaries	Word Accuracy	Word Recall Rate	Word F Value	Sentence Accuracy	Sentence Recall Rate	Sentence F Value
The whole interpretation of Mencius	0.923	0.925	0.924	0.633	0.633	0.633
Annotated Mencius	0.917	0.915	0.916	0.580	0.580	0.580
The proper annotations on Mencius	0.924	0.930	0.926	0.630	0.630	0.630
The combining annotations and commentaries	0.928	0.928	0.928	0.633	0.633	0.633

Of course, using several annotations and commentaries to segment words will also cause some new problems. For example, *The whole interpretation of Mencius*' word list records "後世無", while others only record "後世". "後世無" should be matched according to the maximum matching method, which causes the accuracy loss. We thought that we could add weights such as the word frequency to the words so as to select the best answer in the word list when processing the text by using several annotations and commentaries.

4. Conclusions. This paper mainly studies the automatic word segmentation methods on the *Mencius*. Furthermore, it respectively adopts the Conditional Random Field based Automatic Word Segmentation Method and the Annotation-Commentary Document-based Automatic Word Segmentation Method to segment the words in the *Mencius*, and apart from the word F value, this article also tried to introduce the statistical index, sentence F value, into the experiments. Under the condition of the former method's experiments on the *Mencius*, the maximum word F value has reached 0.944, the maximum sentence F value 0.696. While by using the latter method to conduct the experiments on the *Mencius*, the maximum word F value has arrived at 0.928, the maximum sentence F value 0.633. So to speak, even without model training and machine learning in advance, using annotations and commentaries can gain the result that directly reaches the practical level and reduce the manual labor tremendously, meaning the effect is comparatively prominent.

5. Acknowledgment. This Work is supported by Natural Science Foundation of the Jiangsu Higher Education Institution of China (Grant No.14KJB520023), Humanities and Social Sciences Foundation of The Ministry of Education of China (Grant No.12YJCZH121), Major Projects of Philosophy and Social Science Key Research Bases of Colleges and Universities in Jiangsu Province(Grant No.2010JDXM023).

REFERENCES

[1] Zong Chengqing, Cao Youqi, Yu Shiwen. Sixty Years of Chinese Information Processing. *Applied Linguistics*, Vol.18, No.4, pp.53-61, 2009.

[2] 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.

[3] 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.

[4] Xu Runhua, Chen Xiaohe. A Method of Segmentation on "Zuo Zhuan" by Using Commentaries. *Journal of Chinese Information Processing*, Vol.26, No.2, pp.13-17, 2012.

[5] Shen Dayang, Sun Maosong, Huang Changning. The Implement and the Model of Chinese Segmentation Based on Statistics. *Chinese Information*, No.12, pp. 96-98, 1998(Z1).