A Word Segmentation Method of Ancient Chinese Based on Word Alignment*

Chao Che[1(✉)][0000−0003−2978−5430], Hanyu Zhao[1], Xiaoting Wu[1], Dongsheng Zhou[1][0000−0003−3414−9623], and Qiang Zhang[2]

[1] Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, Dalian University, Dalian 116622, China
{chechao101, hanyuzhao7, wuxiaoting2017}@163.com, donyson@126.com
[2] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
zhangq26@126.com

Abstract. Since there are no public tagged corpora available for ancient Chinese word segmentation (CWS), the state-of-the-art CWS methods cannot be used for ancient Chinese. To address this problem, this paper proposes a word segmentation method based on word alignment (WSWA). Specifically, the method segments words according to the word alignment between modern Chinese words and ancient Chinese characters. If multiple consecutive characters in ancient Chinese align to the same modern Chinese word, they are considered as one word. Because many modern Chinese words are derived from ancient Chinese, the method also exploits the co-occurring characters between modern and ancient Chinese to extract words for CWS. Moreover, to reduce the effect of alignment errors, the method removes the word alignments easily leading to CWS errors. We quantitatively analyze the effects of modern CWS and word alignment on WSWA method using hand-annotated corpora. Our method outperforms the state-of-the-art methods on the WSA experiment on ***Shiji*** with a large margin, which demonstrates the effectiveness of using word alignment to perform ancient CWS.

Keywords: Word Segmentation · Ancient Chinese · Word Alignment.

1 Introduction

Unlike English and other western languages, Chinese words are not delimited by white spaces and CWS is the pre-processing stage of many Chinese nature language processing(NLP) tasks such as information extracting and text mining. Compared to modern Chinese, ancient Chinese is more difficult to segment because it is more concise and compact and has more flexible syntactic structures than modern Chinese. Since the statistical method was applied to CWS in 1990s [1], CWS has made a great progress. Most approaches treat CWS as

* This work is supported by the National Natural Science Foundation of China (No. 61402068)

character sequence labeling problems [2] using character or word features. The state-of-the-art methods can achieve F1 score of around 95% in modern Chinese, depending on what test datasets were used. Nonetheless, few attempts have been made in ancient CWS. To the best of our knowledge, we only found the following research. Shi et al. [3] used the Conditional Random Field (CRF) model to segment some corpora of the pre-Qin period. Qian et al. [4] adopted the Hidden Markov Model (HMM) for CWS on Chuci. Li et al. [5] adapted the capsule architecture to the sequence labeling task to realize Chinese word segmentation for ancient Chinese medical books. They built the tagged ancient Chinese medicine corpora for the word segmentation task. The above work relied on statistical models trained on tagged corpora, which had been built manually for the classic book. However, the construction of tagged corpora is time-consuming and expensive, and there are no public large-scale tagged corpora available for ancient Chinese. To this end, this paper proposes a Word Segmentation method based on Word Alignment (WSWA) to segment ancient Chinese without tagged corpora. The method uses another language with explicit word boundary as the anchor language and performs word segmentation by mapping the word boundary information of the anchor language to ancient Chinese through word alignment. The most common anchor language for CWS is English. However, the bilingual corpus between ancient Chinese and English is rare, and therefore, we regard modern Chinese as a different language from ancient Chinese and take it as the anchor language. Although there are no obvious word delimiters in modern Chinese, the segmentation accuracy of modern Chinese is much higher than that of ancient Chinese. In addition, ancient Chinese and modern Chinese belong to the same language system, and the shared words between them can also be used in word segmentation.

Overall, the main contribution of this paper is as follows:

- WSWA uses a bilingual parallel corpus instead of tagged CWS corpora to solve the problem of lacking large-scale corpora for ancient CWS;
- Taking modern Chinese as the anchor language not only facilitates the acquisition of large-scale bilingual corpora, but also takes advantage of co-occurring characters to extract words;
- Annotation corpora are built manually for modern CWS and word alignment, which are employed to analyze quantitatively the effects of modern CWS and word alignment for ancient CWS.

The rest of this article is organized as follows: Section 2 introduces the related work of word segmentation. WSWA method is detailed in Section 3. In Section 4, we present some experiments and the discussion. Our conclusions are presented in Section 5.

2 Related Work

We divide word segmentation methods into two categories: the monolingual method that only uses the corpus in one language; and the bilingual method that exploits parallel corpora to perform word segmentation.

2.1 Monolingual Word Segmentation

Before 2002, CWS methods were basically based on dictionaries [1] or rules [6]. Since Xue [2], most work have formulated CWS as a sequence labeling task with character tags. Peng et al. [7] first introduced a linear-chain CRFs model to character tagging-based word segmentation. Zhang et al. [8] proposed an alternative, the word-based segmenter, which used a discriminative perceptron learning algorithm and allowed the word-level information to be added as features. Recently, most CWS research focused on neural network. Based on the general neural network architecture for sequence labeling [9], Zheng et al. [10] used character embedding in local windows as input to predict individual character position tags. Following this work, various neural network architectures have been applied to word segmentation, such as max-margin tensor neural network [11], long short-term memory(LSTM) network [12]. Besides sequence labeling schemes, Zhang et al. [13] employed word embedding features for neural network segmentation for transition-based models. Liu et al. [14] proposed a neural segmentation model combining neural network with semi-CRF. Despite of different structures they adopted, the performance of neural segmentation models highly depends on the amount of tagged corpora. Due to the lack of large-scale tagged corpora, most of the above methods cannot be applied to ancient CWS.

2.2 Bilingual Word Segmentation

Word segmentation methods using word alignment can be classified into two lines. One kind of methods utilizes word alignment to extract words from parallel corpora to construct a dictionary, which is then used for word segmentation. The other line makes use of word alignment to refine word segmentation to keep the consistency of segmentation granularity between source and target language so that machine translation can achieve better performance.

For the first line, Xu et al. [15] segmented Chinese words using English words as the anchor language. The Chinese characters are combined into a word if they are aligned with the same English word. Ma and Way [16] also employed the similar idea to do segmentation. Paul et al. [17] learned word segmentation using a parallel corpus by aligning character-wise source language sentences to word units, which was applied to the translation of five Asian languages into English. For the other line, Wang et al. [18] explored the use of a manually annotated word alignment corpus to refine word segmentation for machine translation. The words were aligned to minimum translation unit, which was English words plus the compounds. Tran et al. [19] proposed a new method to re-segment words in both Chinese and Vietnamese in order to strengthen 1-1 alignments and enhance machine translation performance. They adjusted WS in both Chinese and Vietnamese based on four factors, namely NE, Sino-Vietnamese shared language, word level alignment result, and character-word level alignment result.

The bilingual word segmentation usually leverages a language with explicit word boundary markers as anchor language. However, a language without obvious delimiters but has a segmenter tool of high performance can also be

treated as anchor language. For example, Chu et al. [20] refined CWS based on Chinese-Japanese parallel corpora. Japanese does not have white spaces between words. However, a Japanese segmenter toolkit can have F1 score of up to 99%. Moreover, they also exploited common Chinese characters shared between Chinese and Japanese in CWS optimization. Our WSWA method also leverages the shared characters between ancient and modern Chinese by extracting co-occurring words to perform CWS.

3 WSWA

3.1 Monolingual Word Segmentation

The WSWA method performs ancient CWS based on the following two ideas. On one hand, since modern Chinese has been derived from ancient Chinese, some lexicon information of ancient Chinese such as person and official names have been reserved in modern Chinese. If many consecutive characters co-occur in both ancient and modern Chinese, they are most likely to be words kept from the ancient times, and should be regarded as a word. On the other hand, modern Chinese has very good performance for WS due to abundant language resources. We can leverage word boundary information in modern Chinese by mapping the characters of ancient Chinese to those of modern Chinese through word alignment. If more than one characters in ancient Chinese align to the same word in modern Chinese, the characters express the same meaning and should be merged into a word. The idea can be explained through an example shown in Fig. 1. In Fig. 1, the ancient Chinese sentence " 庄襄王为秦质子于赵" corresponds to " 庄襄王在赵国作秦国人质时" in modern Chinese, and the alignment between the characters of the ancient Chinese sentence and the words of the modern Chinese sentence is shown. " 庄襄王" appears in both sides, it is extracted as a word, first. " 质", " 子" are aligned to the same word " 人质", so the two characters can be combined as a word. " 为" matches the word " 作", so we treat it as a single word. Similarly, " 秦", " 于" and " 赵" align to " 秦国", " 在" and " 赵国", respectively. They are all separated as single words. Finally, the ancient Chinese can be segmented as "庄襄王/为/秦/质子/于/赵" according to the character co-occurrence and the word alignment relationship between ancient Chinese and modern Chinese.

Given the parallel corpus between modern and ancient Chinese, WSWA method segments ancient Chinese in the following steps:

- Step 1: Divide ancient Chinese into single characters and segment modern Chinese into words with parts of speech.
- Step 2: Extract co-occurring characters in ancient Chinese. If several consecutive ancient Chinese characters also appear in modern Chinese, they are extracted as a word from ancient Chinese.
- Step 3: Employ IBM-3 model to implement alignment between modern Chinese words and ancient Chinese characters.

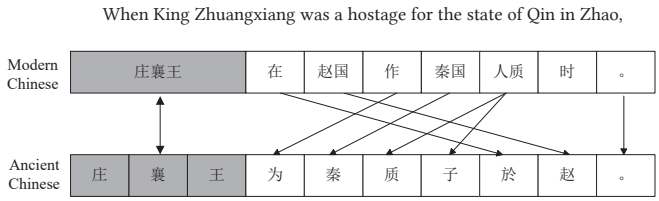

Fig. 1: A word segmentation example.

- Step 4: Remove word alignment errors and the word alignments in the deleting list.
- Step 5: Merge ancient Chinese characters into words according the word alignment. If multiple consecutive characters correspond to one modern Chinese word, combine them into a word. In addition, characters representing numbers are combined into a word.

Next, we will explain some problems in word alignment of step 3 in section 3.2 and give a detailed introduction of step 4 in section 3.3.

3.2 Word Alignment

In WSWA, we employ IBM-3 model [21], which is implemented by GIAZ++ [3], to perform alignment between modern Chinese words and ancient Chinese characters. IBM-3 has a limitation in modeling n-1 alignment, but it will not affect our segmentation results. IBM-3 treats n-1 alignment as several 1-0 alignments and one 1-1 alignment. Specifically, the alignment many modern Chinese words corresponds to one ancient Chinese character will be handled as several alignments that one modern Chinese words maps to null and one alignment that one modern Chinese words map to one ancient character. The WSWA combines the characters aligned to the same modern Chinese word as a word. So 1-0 alignments have no effect on our segmentation result, the ancient Chinese character in 1-1 alignment can also correctly segmented as a word.

Although two sides of alignment are from the same language, we treat the alignment as a bilingual problem instead of monolingual alignment. Ancient Chinese and modern Chinese are more like two different languages due to huge lexical and syntax difference. For example, “妻子” means wife and children in ancient Chinese while it only refers to wife in modern Chinese. Thus, monolingual alignment, which mainly exploiting word similarity and contextual evidence to discover and align similar semantic units in a natural language, cannot work well in the alignment between modern and ancient Chinese.

[3] https://codeload.github.com/moses-smt/giza-pp/zip/master

3.3 Deleting Alignment Errors

Because of the performance limitation of current alignment method, there are some errors in the alignment, such as the alignment between punctuation and characters. The alignment errors will result in word segmentation failure, since words are segmented using the matching relationship in word alignment. To reduce the segmentation errors caused by incorrect alignment, we process the alignment as follows:

- Remove the alignment errors. Low alignment probability denotes inaccurate alignment, therefore we remove the alignment with probability less than a very small threshold (0.0001). The alignment, in which an ancient character corresponds to non-Chinese character such as punctuation, is also deleted for they are obviously incorrect.
- Remove the alignment that easily leads to segmentation errors. Some function words in ancient Chinese usually align to null in modern Chinese. Those words often cause alignment errors since they frequently appear after some nouns. Thus, we collect those words in a deleting list and remove them from word alignments in case they cause segmentation errors. The deleting list contain 16 words, namely, ' 乎', ' 也', ' 以', ' 乃', ' 亦', ' 立', ' 曰', ' 遂', ' 已', ' 尔', ' 矣', ' 则', ' 在', ' 哉', ' 悉', ' 而'. Most words in deleting list are function words and some function words have multiple parts of speech. Hence, deleting them directly will cause some segmentation errors for some person names and places. For example, " 耳" is commonly used as a function word in ancient Chinese and needn't to be translated, but it also appears in some person names such as " 重耳" and " 张耳". To eliminate the influence of the function words and not make more mistakes, we first determine the parts of speech of the words in the deleting list according to the modern Chinese then remove the words whose parts of speech are function word.

4 Experiments

4.1 Experiment Settings

The parallel corpus used in the experiments includes five basic annals from Shiji: Annals of Qin, the Basic Annals of the First Emperor of the Qin, the Basic Annals of Hsiang Yu, the Basic Annals of Emperor Kao-tsu and the Basic Annals of Empress Li. The corpus contains 4145 sentence pairs of ancient and modern Chinese. In the corpus, the vocabulary size of ancient Chinese is 4285 and modern Chinese has 6429 words.

The evaluation measure of word segmentation: The experimental results of word segmentation were measured by precision (P), recall (R), and $F1$ measure, whose definition can be seen in [22].

The evaluation measure of word alignment: We used Alignment Error Rate (AER) to evaluate word alignments. Given that the alignment result de-

noted as set A, and the gold standard manually aligned result denoted as set S, AER can be defined as:

$$AER = 1 - \frac{2|A \cap S|}{|A| + |S|} \tag{1}$$

4.2 Experiment Result

Our experiments consists of four parts. In the first part, we validated the effectiveness of our segmentation idea and different measures we proposed to reduce the segmentation errors. In the second and third part, we analyzed the impact of modern CWS and word alignment on ancient CWS, respectively. In the last part, we compared WSWA method with state-of-the-art monolingual segmentation methods on ancient CWS.

The analysis of different processing measures: To investigate the upper bound performance of WSWA in theory, we performed WSWA on hand-tagged modern CWS corpus and hand-tagged word alignment corpus. The comparison results are shown in Table 1. The use of hand-tagged corpora isolates the influence of modern CWS errors and word alignment errors on the ancient CWS. From Table 1, we can see that the WSWA method can achieve F1 as high as 99.1%, which confirms the effectiveness of WSWA. However, there still exists 0.9% segmentation errors caused by omitting words. For the fluency of translation, we omit some ancient Chinese words when translating into modern Chinese, which will lead to word segmentation errors in ancient Chinese. For example, in sentence “於是項王乃欲东渡乌江” which means “At this point, King Hsiang had intended to cross east over [the Yangtze River] from Wu-chiang.” , “於是” is not translated in the modern Chinese “项王想要向东渡过乌江” . Therefore, “於” and “是” in ancient Chinese cannot be combined correctly because there are no alignment words for them.

In this paper, we improve CWS accuracy by extracting co-occurring characters and deleting the alignment errors. To validate whether the measure works, we ran WSWA methods with no measure, only one single measure and all the measures, respectively. The results are shown in Table 1. All the WSWA methods employed NLPIR [4] for modern CWS and implemented GIZA++ for word alignment.

As can be seen Table 1, extracting co-occurring characters is a simple but very effective method. It significantly improves the performance of word segmentation especially for recall. Because of the wide distribution of co-occurring characters in both ancient and modern Chinese, the measure can extract the words not combined by word alignment.

At the same time, this measure is implemented by word alignment and can reduce some word alignment mistakes. When a character appears more than one time in one ancient Chinese sentence, it is very difficult to get the right alignment. Taking sentence pair “立二世之兄子公子婴为秦王” and “就立二世

[4] http://ictclas.nlpir.org/

Table 1: The results comparison of WSWA methods using different measures.

Extracting Co-occurring Characters	Deleting Alignment Errors	P	R	F1
		80.3%	65.9%	72.4%
+		81.7%	**85.0%**	83.3%
	+	81.1%	67.6%	73.8%
+	+	**86.9%**	81.6%	**84.2%**
WSWA method using hand-annotated corpora		**99.1%**	**99.1%**	**99.1%**

哥哥的儿子公子嬰为秦王" (Then he set up Ziying, the son of one of the Second Emperor' s older brothers, as king of Qin.) for example, character "子" appears twice in ancient Chinese, the candidate alignment words "儿子", "公子", "公子嬰" of the character "子" all occur in the same modern Chinese sentence. It is hard to determine which word each "子" should align to. After extracting co-occurring characters "公子嬰", there is only one candidate word "儿子" left. It is easy to find that the first "子" should map to "儿子".

Table 1 shows that the performance of our method can also be boosted by deleting alignment errors. Deleting alignment errors reduces the word segmentation errors propagating from word alignment. For example, the sentence " 申侯之女" is segmented wrongly as " 申侯之/女" if not deleting alignment errors. The function word " 之" is easily aligned wrongly to the noun " 申侯" before it, since it often appears after the nouns. After deleting the function words, we get the right segmentation result " 申侯/之/女".

The influence of modern CWS: We employed three state-of-the-art methods to perform modern CWS, namely Jieba [5], Stanford [6] and NLPIR. The performance of three methods on modern Chinese are listed in the column "modern CWS" of Table 2. Based on the modern CWS results of three methods, we performed WSWA method on ancient Chinese three times, respectively, the performance of which is listed in the column "WSWA" of Table 2. To avoid the impact of alignment errors, word alignment between ancient and modern Chinese is labeled by hand.

Table 2: The performance of WSWA using different modern CWS methods.

Segmentation Methods	Modern CWS			WSWA		
	P	R	F1	P	R	F1
Jieba	80.5%	**86.2%**	83.3%	90.4%	**90.4%**	90.4%
Stanford	82.9%	85.8%	84.3%	91.7%	90.0%	90.8%
NLPIR	**89.5%**	81.7%	**85.4%**	**92.2%**	90.0%	**91.1%**

[5] https://github.com/fxsjy/jieba
[6] http://nlp.stanford.edu/software/segmenter.shtml

Table 2 shows that the word segmentation result of modern Chinese has direct impact on those of ancient Chinese. The method that used a better modern CWS result always performed better on ancient CWS. The reason is simple and straight forward. If the word boundary information in modern Chinese is wrong, the information transferring to ancient Chinese is also wrong.

In Table 2, it is noted that the performance of WSWA based on modern CWS results is much higher than that of modern CWS, which is counterintuitive. The ancient Chinese word is concise, and the modern Chinese explanation adds a lot of words in order to make the sentence fluent. Many modern Chinese words that segmented incorrectly do not appear in ancient Chinese, so they do not affect the WS results of ancient Chinese.

The three methods have similar performance on modern CWS. Since NLPIR outperformed the other two methods on precision and WSWA using the result of NLPIR has the best performance, we selected NLPIR for modern CWS in our test.

The influence of word alignment: We conducted alignment between ancient Chinese characters and modern Chinese words by two alignment tools, GIZA++ and BerkeleyAligner [7], which implement IBM-3 model and bidirectional HMM model, respectively. The AER of two alignment models is shown as Table 3. We also ran WSWA using the alignment of two models on the hand-tagged modern CWS corpora to test the influence of alignment to ancient CWS. The results are also shown in Table 3.

Table 3: The performance of WSWA using different word alignment models.

Alignment models	AER	WSWA		
		P	R	F1
IBM-3	0.128	89.9%	**93.5%**	91.7%
HMM	0.377	**91.3%**	89.5%	90.4%

From Table3, it is clear that the performance of the alignment model is closely related to the ancient CWS results. The IBM-3 model with lower AER outperforms HMM model for ancient CWS because word alignment errors will lead to the wrong combination of characters. For instance, in the sentence “与晋战河阳” (He fought with the state of Jin at Heyang) and its modern Chinese translation “和晋国交战于河阳”, “晋” should map to “晋国”. If “晋” and “战” are aligned wrongly to “交战”, they are combined as a word “晋战” by mistake. However, word alignment errors not always result in wrong word alignment and we can obtain correct words based on wrong word alignment sometimes. For example, in sentence “善哉乎贾生推言之也” (Master Jia has written an excellent discussion of the matter.) and its translation “贾生论述的非常好”, instead of aligning “推” and “言” to “论述”, we map both of

[7] https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/berkeleyalignerss

them to “非常” wrongly. Nevertheless, they can be merged as a word “推言” correctly because they all align to the same word.

Comparison with monolingual word segmentation method: To test the effectiveness of WSWA method, we compared WSWA method with three monolingual word segmentation methods, i.e., Jieba, Stanford and NLPIR with WSWA method on CWS. We first employed the three methods to directly segment all the 4145 ancient Chinese sentences, which are already trained on modern Chinese corpus. The result is shown as Table 4. For the sake of fairness, we also retrained Jieba and Stanford segmenter on ancient Chinese corpus and then ran ancient CWS test, as shown in Table 5. We divided the training set and the test set and at the rate of 4:1. Specifically, the WS methods were trained on 3316 sentences and were tested on 829 sentences.

Table 4: Result comparison with word segmentation methods trained on modern Chinese corpus.

Segmentation Methods	P	R	F1
Jieba	56.0%	69.1%	61.9%
Stanford	60.7%	72.7%	66.1%
NLPIR	80.2%	76.9%	78.5%
WSWA	**87.4%**	**81.9%**	**84.5%**

Table 4 shows that the untrained Jieba and Stanford segmenters have poor performance on ancient Chinese. There is a big performance gap between them and WSWA method. This confirms the conclusion that the model trained on modern Chinese cannot be applied to ancient Chinese due to the great syntax and grammar difference between them. NLPIR outperformed other two methods obviously. Considering they have similar performance on modern CWS, we guess the dictionary of NLPIR may include many ancient Chinese words.

Table 5: Result comparison with word segmentation methods re-trained on ancient Chinese corpus.

Segmentation Methods	P	R	F1
Jieba	58.0%	70.1%	63.5%
Stanford	78.5%	63.2%	70.0%
WSWA	**84.6%**	**77.8%**	**81.0%**

NLPIR is not re-trained because it does not provide API for retraining. In Table 5, Jieba and Stanford segmenters only gain a small performance enhancement after re-trained on ancient Chinese since the statistical methods need to be

trained on large-scale tagged corpus while small-scale data cannot let them learn the features fully. Our method outperformed them with a large gap in precision, recall and F1 since our method leverages the word boundary information from modern Chinese by word alignment, thus it is not affected by the scale of tagged corpora.

5 Conclusions

In this paper, we proposed WSWA method to perform ancient CWS without tagged corpus. WSWA method segmented ancient Chinese by transferring the word boundary information from modern Chinese to ancient Chinese through word alignment. In the experiment on ***Shiji***, WSWA outperformed other segmentation methods. Although the precision of WSWA is far from that of the state-of-the-art word segmentation methods trained on large-scale corpora, WSWA has relatively high precision on proper nouns, which makes it suitable for the NLP tasks which focus on terms such as term alignment and name entity recognition. As for other NLP tasks, the word segmentation should be refined by other methods.

The performance of WSWA method highly depends on the quality of word alignment and modern CWS. Therefore, we will try to reduce the errors in word alignment and modern CWS to further enhance the performance of word segmentation method in the future.

References

1. Sproat, R., Shih, C., Gale, W., Chang, N.: A stochasitic finite-state word-segmentation algorithm for Chinese. Computational Linguistics **22**(3), 377-404 (1996)
2. Xue, N., Shen, L.: Chinese word segmentation as LMR tagging. In: Sighan Workshop on Chinese Language Processing. pp. 176-179 ACL, Stroudsburg (2003)
3. Shi, M., Bin, L.I., Chen, X.: CRF based research on a unified approach to word segmentation and POS tagging for pre-qin Chinese. Journal of Chinese Information Processing (2010)
4. Qian, Z., Zhou, J., Tong, G., Su, X.: Research on automatic word segmentation and POS tagging for Chu Ci based on HMM. Library and Information Service **58**(4), 105-110 (2014)
5. Li, S., Li, M.Z., Xu, Y.J., Bao, Z.Y., Fu, L., Zhu, Y.: Capsules Based Chinese Word Segmentation for Ancient Chinese Medical Books. IEEE Access **6**, 70874–70883 (2018)
6. Palmer, D.D.: A trainable rule-based algorithm for word segmentation. In: Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. pp. 321-328 ACL, Stroudsburg (1997)
7. Peng, F., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. Proceedings of Coling. pp. 562-568 (2004)

8. Zhang, Y., Clark, S.: Chinese segmentation with a word-based perceptron algorithm. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 840-847 ACL, Stroudsburg (2007)
9. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**(1), 2493-2537 (2011)
10. Zheng, X., Chen, H., Xu, T.: Deep learning for Chinese word segmentation and POS tagging. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 647-657 ACL, Stroudsburg (2003)
11. Pei, W., Ge, T., Chang, B.: Max-margin tensor neural network for Chinese word segmentation. In: In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. pp. 293-303 ACL, Stroudsburg (2014)
12. Chen, X., Qiu, X., Zhu, C., Pengfei, L., Huang, X.: Long short-term memory neural networks for chinese word segmentation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1197-1206 ACL, Stroudsburg (2015b)
13. Zhang, M., Zhang, Y., Fu, G.: Transition-based neural word segmentation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 421-431 ACL, Stroudsburg (2016)
14. Liu, Y., Che, W., Guo, J., Qin, B., Liu, T.: Exploring segment representations for neural segmentation models. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 2880-2886 (2016)
15. Xu, J., Zens, R., Ney, H.: Do we need Chinese word segmentation for statistical machine translation? In: ACL SIGHAN Workshop Association for Computational Linguistics. pp. 122-128 ACL, Stroudsburg (2004)
16. Ma, Y., Way, A.: Bilingually motivated domain-adapted word segmentation for statistical machine translation. In: Conference of the European Chapter of the Association for Computational Linguistics. pp. 549-557 ACL, Stroudsburg (2009)
17. Paul, M., Finch, A.M., Sumita, E.: Integration of multiple bilingually-trained Segmentation Schemes into Statistical machine translation. In: Joint Fifth Workshop on Statistical Machine Translation and Metricsmatr. pp. 400-408 (2010)
18. Wang, X., Utiyama, M., M. Finch, A., Sumita, E.: Refining word segmentation using a manually aligned corpus for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1654-1664 ACL, Stroudsburg (2014)
19. Tran, P., Dinh, D., Nguyen, L.H.B.: Word re-segmentation in Chinese-Vietnamese machine translation. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) **16**(2), 12 ACM, New York (2016)
20. Chu, C., Nakazawa, T., Kawahara, D., Kurohashi, S.: Chinese-Japanese machine translation exploiting Chinese characters. ACM Transactions on Asian Language Information Processing **12**(4), 1-25 ACM, New York (2013)
21. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J.: The mathesmatics of statistical machine translation: parameter estimation. Computational Linguistics pp. 263-311 (1993)
22. Wu, X.T., Zhao, H.Y., Che, C.: Term Translation Extraction from Historical Classics Using Modern Chinese Explanation. In: The 17th China National Conference on Computational Linguistics & 6th International Symposium on Natural Language Processing Based on Naturally Annotated Big Data (CCL 2018/NLP-NABD 2018). pp. 88–98 CCL, Beijing (2018)