Exploring Multiple Chinese Word Segmentation Results Based on Linear Model

Chen Su, Yujie Zhang, Zhen Guo, and Jinan Xu

School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 100044, China
{12120447,yjzhang,12120416,jaxu}@bjtu.edu.cn

Abstract. In the process of developing a domain-specific Chinese-English machine translation system, the accuracy of Chinese word segmentation on large amounts of training text often decreases because of unknown words. The lack of domain-specific annotated corpus makes supervised learning approaches unable to adapt to a target domain. This problem results in many errors in translation knowledge extraction and therefore seriously lowers translation quality. To solve the domain adaptation problem, we implement Chinese word segmentation by exploring n-gram statistical features in large Chinese raw corpus and bilingually motivated Chinese word segmentation, respectively. Moreover, we propose a method of combining multiple Chinese word segmentation results based on linear model to augment domain adaptation. For evaluation, we conduct experiments of Chinese word segmentation and Chinese-English machine translation using the data of NTCIR-10 Chinese-English patent task. The experimental results showed that the proposed method achieves improvements in both F-measure of the Chinese word segmentation and BLEU score of the Chinese-English statistical machine translation system.

Keywords: Chinese Word Segmentation, Domain Adaptation, Bilingual Motivation, Linear Model, Machine Translation.

1 Introduction

In the process of developing domain-specific Chinese-English machine translation (MT) system, the accuracy of Chinese word segmentation on large amount of training text often decreases because of unknown words. The lack of domain-specific annotated corpus makes supervised learning approaches unable to adapt to a target domain. When extracting translation knowledge from large-scale Chinese-English parallel corpus, the poor accuracy of Chinese word segmentation results in many errors in extracted translation knowledge and therefore seriously lowers translation quality.

To solve this problem, a few approaches have been applied to domain adaptation of Chinese word segmentation. (Zhang et al., 2012) indicates that adding domain-specific dictionary to CRF-based Chinese word segmentation is effective for domain adaptation. In cases where no domain-specific dictionaries are available, n-gram statistical features are explored from large Chinese raw corpus to replace dictionary

G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 50–59, 2013.

(Wang et al., 2011) (Guo et al., 2012). When Chinese-English parallel corpus is given for developing MT system, words in English sentences may guide Chinese word segmentation of the corresponding Chinese sentences (Ma et al., 2009) (Xi et al., 2012). (Ma et al., 2010) shows that combining different Chinese word segmentation results can improve the performance of statistical machine translation system.

In this paper, we make improvements and extensions to the above approaches and implement two Chinese word segmenters, one based on n-gram features of Chinese raw corpus and the other one based on bilingually motivated features. Furthermore, we propose a method, by which the multiple Chinese word segmentation results of the two segmenters are combined based on linear model to augment domain adaptation. In this way, we obtain an adapted Chinese word segmenter for a large scale domain-specific Chinese text.

The rest of this paper is organized as follows. In Section 2, we introduce the implementation of the two Chinese word segmenters and then describe the method of combining multiple Chinese word segmentation results in detail. The evaluation experiments conducted in patent domain are reported in Section 3. Section 4 concludes and gives avenues for future work.

2 Combining Multiple Segmentation Results with Linear Model

In the process of developing domain-specific Chinese-English machine translation system, a large-scale parallel Chinese-English corpus needs to be processed for translation knowledge extraction. There also exists a large-scale Chinese raw corpus of the same domain that needs to be translated by the MT system. For exploring these resources to improve the accuracy of Chinese word segmentation, we attempt to take n-gram features of the Chinese raw corpus and bilingually motivated features of the parallel corpus into account. The idea will be explained in more detail as follows. N-gram statistical features of the domain-specific Chinese raw corpus are to be explored for adapting a Chinese word segmenter to a specific domain. And bilingually motivated features of the parallel corpus are to be utilized as guidance for segmenting domain-specific Chinese words. In order to integrate the two kinds of features from different types of corpora, we implement two Chinese word segmenters based on n-gram features and bilingually motivated features respectively and then adopt a linear model to combine the results of the two segmenters. The framework of the idea is shown in Fig. 1.

In this paper, we implement a Chinese word segmenter based on CRF model, in which n-gram statistical features of Chinese raw corpus are integrated, called as CRF segmenter in this paper. Moreover, for using bilingually motivated features, we implement a Chinese word segmentation system based on word alignment techniques, called as bilingually motivated segmenter in this paper. We will describe the two segmenters in the following two subsections respectively, and give a detailed description about how to combine their results in the third subsection.

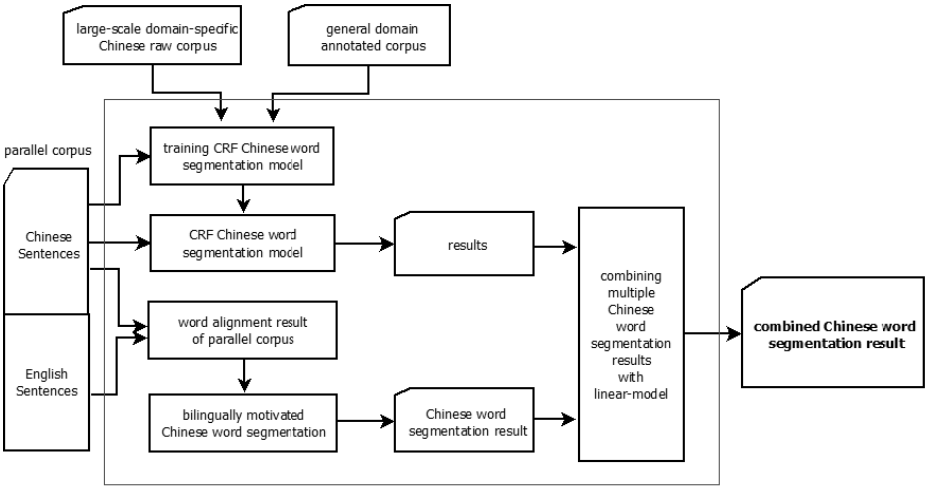

Fig. 1. Framework of combining multiple Chinese word segmentation results

2.1 Domain-Adapted Chinese Word Segmentation Based on N-gram Features

Following the work of (Guo et al., 2012), we realize a domain-adapted Chinese segmentation system by exploring Chinese raw corpus of target domain.

In additional to the UPENN Chinese annotated data[1], statistical features of large-scale domain-specific Chinese raw corpus are added for training segmentation model. The overview of the domain adaptation for Chinese Word Segmentation is shown in Fig. 2. We use CRF++ (version 0.55)[2] to train segmentation model. In this paper, n-gram refers to a sequence of n consecutive Chinese characters. The statistical features of n-gram include n-gram frequency feature and AV (Accessor Variety) feature (Feng et al., 2004), defined as the count of occurrences of n-gram in a corpus and the count of different context in which n-gram occurs, respectively. The feature template adopts 5-characters sliding window: two preceding characters, a current character and two subsequent characters (Jin et al., 2005).

For a sentence, CRF segmentation model produces an N-best list of alternative segmentation results with corresponding probability scores. In previous work, only 1-best result is adopted generally. By analyzing the results within N-best, however, we find that some erroneous segmentation in 1-best result may be segmented correctly in the low-ranking results, such as the example shown in Fig. 3. In this example, the character sequence "甘氨酸"(Glycine) is wrongly segmented into two words in 1-best result, whereas it is correctly segmented into one word in the 3-best result. Based on the observation, we intend to select correctly segmented parts from the N-best list. In this paper we use a 10-best list and denote the corresponding probability scores as $Conf_{CRF1}$, $Conf_{CRF2}$, …, $Conf_{CRF10}$, which will be used to measure the confidence of words in the k-best result ($1 \leq k \leq 10$) (see Section 2.3).

[1] http://www.ldc.upenn.edu/Catalog/
[2] https://code.google.com/p/crfpp/

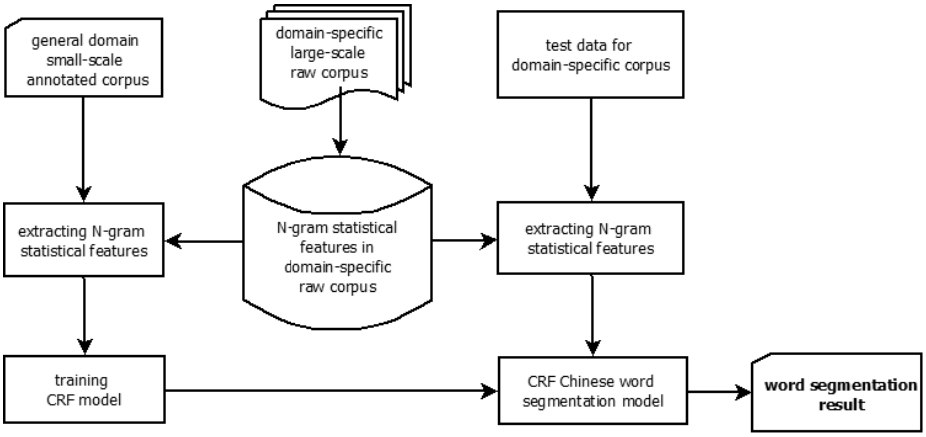

Fig. 2. Framework of domain adaptation for a Chinese word segmentation system

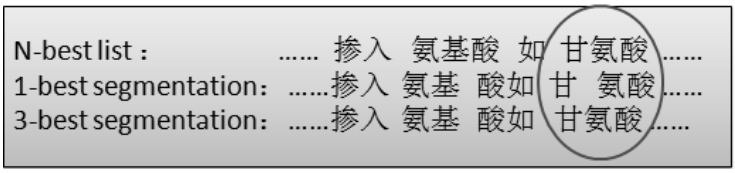

Fig. 3. An example of a correct segmentation occurring in 3-best result

2.2 Bilingually Motivated Chinese Word Segmentation

In Chinese-English parallel corpus, word boundaries in English sentences are orthographically marked. According to alignment between words of an English sentence and Chinese characters of the corresponding Chinese sentence, the boundaries of Chinese word are inferable. This can be illustrated by the example shown in Fig. 4. In Fig. 4, a part of alignment results displayed as lines between the English sentence and the Chinese sentence is shown. English words, "sebacic", "acid-derived" and "monomer" are aligned with Chinese character sequence, "癸二", "酸衍生" and "单体", respectively. The alignment results imply that the Chinese characters sequence "癸二酸衍生单体" is likely to be segmented into "癸二", "酸衍生" and "单体". So the aligner between English words and Chinese characters may guide Chinese word segmentation, especially for domain-specific words. Based on the above consideration, we implement a bilingually motivated Chinese word segmentation using word alignment results.

Given a Chinese-English sentence pair, $C_1^J = c_1 c_2 \ldots c_J$ and $E_1^I = e_1 e_2 \ldots e_I$, in which c_j ($1 \le j \le J$) is the jth character in the Chinese sentence, and e_i ($1 \le i \le I$) is the ith word in the English sentence. We use GIZA++ toolkit[3] for word alignment. It should be noted that Chinese characters in Chinese side are used as units for

[3] https://code.google.com/p/giza-pp/

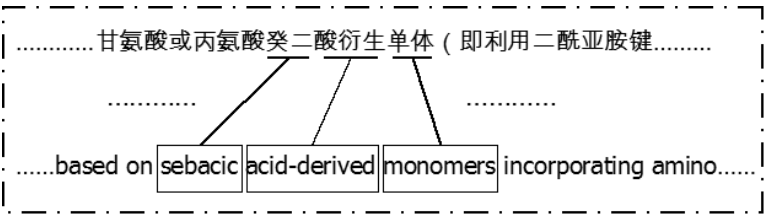

Fig. 4. An example of bilingually motivated Chinese word segmentation

alignment. Two directions of alignment processing, i.e. from Chinese to English and from English to Chinese, are conducted respectively. After that, the alignment results of the two directions are integrated by using heuristic grow-diag-final[4]. Let a_i=< e_i, C> represent an alignment from one single English word e_i to a set of Chinese characters C. When the Chinese characters of C are consecutive in the Chinese sentence, C is regarded as a Chinese word.

Let *Count* (a_i) represents the number of the alignment a_i occurring in the alignment results and *Count* (e_i, C) represents the co-occurrence frequency, i.e. the number of times C and e_i co-occur in the parallel corpus. Let *Conf* (a_i) represents the confidence score of a_i, by which we measure the possibility of e_i being aligned with C when they co-occur in one sentence pair. The value of *Conf*(a_i) is therefore estimated as in (1).

$$Conf(a_i) = \frac{Count(a_i)}{Count(e_i, C)} \tag{1}$$

The algorithm of the bilingually motivated Chinese word segmentation is defined as follows.

(a) Conduct word alignments in two directions by using GIZA++ and integrate the alignment results. In using GIZA++, each character of Chinese sentence is regarded as one word.
(b) Calculate confidence scores of alignment results according to (1).
(c) For each alignment a_i=< e_i, C>, take C as a word segmentation and the confidence score of the word as *Conf*(a_i), if the characters in C are consecutive in the sentence.

2.3 Combining Multiple Segmentations with Linear Model

We will describe the method of combining the multiple segmentation results of the two segmenters, which have been presented above.

The n-gram features are extracted from Chinese monolingual corpus, while the bilingually motivated features are extracted from bilingual corpus. The two types of features belong to different kinds of language resources. In order to integrate the different kinds of features in Chinese word segmentation, we instead integrate the results from the different segmenters which are implemented using different type of features

[4] http://www.statmt.org/moses/?n=FactoredTraining.AlignWords

respectively. For this purpose, we design a linear model to combine the multiple segmentation results for selecting the best Chinese word segmentation.

We label boundaries between Chinese characters in a sentence with sequential numbers. Such an example is shown in Figure 5, in which the numbered boundaries are displayed as nodes. B_i^j denotes the Chinese character sequence between the *i*th and the *j*th boundaries in the sentence.

From the CRF segmenter, we adopt the segmentation results in a 10-best list and take the probability score $Conf_{CRFk}$ as the confidence score of corresponding words in the *k*-best result ($1 \le k \le 10$). We then adopt the segmentation result of the bilingually motivated segmenter as the eleventh segmentation result and the confidence score of alignment $Conf(a_i)$ is taken as the confidence score of corresponding words in the result. We designed a linear model as in (2) to combine the 11 segmentation results.

$$F_{i,j} = \lambda_1 \cdot Conf_{CRF1} \cdot seg_1(i,j) + \cdots + \lambda_{10} \cdot Conf_{CRF10} \cdot seg_{10}(i,j) + \lambda_{11} \cdot Conf_{i,j} \cdot seg_{11}(i,j) \quad (2)$$

Where $F_{i,j}$ denotes the possibility score of B_i^j being a word; $seg_l(i, j)$ ($1 \le l \le 11$) is a two-valued function, $seg_l(i, j)=1$ when B_i^j being a word in the *l*th segmentation result; otherwise, $seg_l(i, j)=0$; $Conf_{CRFk}$ ($1 \le k \le 10$) is the confidence score of the *k*th segmentation result from the CRF segmenter; $Conf_{i,j}$ is the confidence score of the segmentation result from the bilingually motivated segmenter; λ_l ($1 \le l \le 11$) are weights of 11 segmentation results.

$$w_{i,j} = \frac{F_{i,j}}{\sum_j F_{i,j}} \quad (3)$$

$F_{i,j}$ is normalized into $w_{i,j}$ as in (3). $w_{i,j}$ is the normalized score of a word. We then represent the multiple candidate words in a lattice, as shown in Fig. 5. The nodes marked with numbers represent boundaries between characters and the directed edge <*i*, *j*>, from nodes *i* to node *j*, represents that B_i^j is a word with a normalized score $w_{i,j}$. The best segmentation result should be a sequence of words with a maximum product of their scores. Such a sequence can be found by a dynamic programming, also called decoding for lattice.

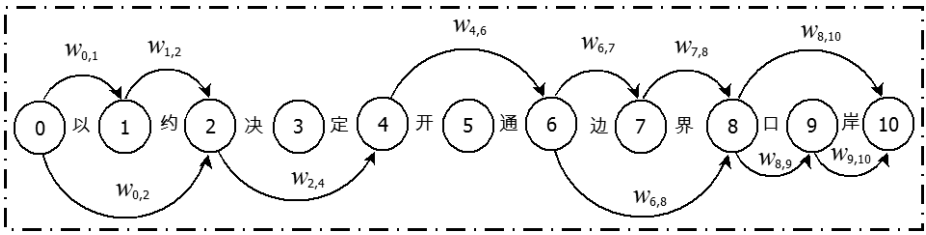

Fig. 5. An example of Chinese word segmentation lattice

To train the parameters λ_l ($1 \le l \le 11$), we use Powells algorithm combined with a grid-based linear optimization method as follows (William et al., 2002). First, a point in 11-dimensional parameter space is randomly selected as a initial point and then the parameters λ_l are optimized through iterative process. In each step, only one

parameter is optimized w.r.t. F-measure of word segmentation, while keeping all other parameters fixed. To avoid local optimum, we select different starting points for parameter estimation.

In this way, we obtain a linear model, by which the best word segmentation is selected from the results of the two domain-adapted segmenters and therefore domain-adaptation of Chinese word segmentation is effectively augmented.

3 Experimental Result

For verifying the contribution of the proposed method, we consider a practical case of machine translation system development. For this aim, the experiments are designed on the NTCIR-10[5] Chinese-English parallel patent description sentences. We carry out word segmentation on the Chinese part of the data by using the proposed method and conduct evaluations from two aspects, accuracy of Chinese word segmentation and translation quality of MT system.

3.1 Experimental Data

The data of NTCIR-10 provides 1,000,000 sentence pairs for training, 2,000 sentence pairs for development and 2,000 sentence pairs for testing. From the training set, 300 sentence pairs are randomly selected as annotation set, denoted as AS. The remaining sentences are used as training set, denoted as TS. The purpose of this work is to increase the accuracy of Chinese word segmentation on the Chinese part of TS and bring about improvement on translation quality of MT system. We annotate the Chinese sentences of AS according to the word segmentation specification of the Penn Chinese Treebank (Xia et al., 2000). The annotated 300 sentences are then randomly divided into two parts, denoted as AS1 and AS2, used for training the parameters of the linear model and evaluating Chinese word segmentation accuracy, respectively.

The Penn Chinese Treebank (CTB 5.0), including chapter 1-270, chapter 400-931 and chapter 1001-1151, are used as annotated corpus to train the CRF segmentation model.

3.2 Evaluation of Chinese Word Segmentation

To train the domain-adapted CRF segmentation model, we use CTB data as annotated data and take the Chinese sentences of TS as a large raw corpus for extracting the n-gram features. The CRF segmentation model adapted to the patent domain is then used to segment the Chinese sentences of TS and the segmentation results of 10-best list for each sentence are kept as candidate words.

We also conduct word alignment on the Chinese-English parallel sentence pairs of TS and obtain bilingually motivated Chinese word segmentation results on the Chinese sentences of TS according to the algorithm described in Section 2.2.

[5] http://research.nii.ac.jp/ntcir/ntcir-10/

The parameters λ_l ($1 \le l \le 11$) of the linear model are trained by the 150 annotated sentences of AS1. At last, we use the linear model to combine the 11 Chinese segmentation results and obtain the best word segmentation results on the Chinese sentences of TS.

To evaluate the accuracy of the Chinese word segmentation, we segment the 150 sentences of AS2 in the same way and show the evaluation results in Table 1. We can see that the linear model outperforms the other two methods. Compared with the 1-best of CRF segmenter, the accuracy of the linear model is increased by 1% both on precision and on recall rate, and F-measure is therefore increased by 1.257%. It proves that the proposed method of combining multiple word segmentation results yields higher performance in adapting the Chinese word segmentation to the patent domain. We investigate the segmented sentences of TS and find that there are totally 37,109,126 Chinese words. This implies that the slight variation in the accuracy of Chinese word segmentation is of practical value. Furthermore, an improvement in translation quality of MT system is expected to be achieved by the improvement of Chinese word segmentation.

Table 1. Comparison of the evaluation results of three domain-adapted Chinese word segmentation methods

Chinese word segmentation method	Precision[%]	recall[%]	F-measure[%]
Bilingually motivated segmenter	73.1312	61.4480	66.7825
1-best of CRF segmenter	90.2439	90.7710	90.5067
Our method(Combining multiple segmentation results)	91.6650	91.8614	91.7631

3.3 Evaluation of Machine Translation System

We then develop a phrase-based statistical machine translation system on the Chinese-English parallel sentences of TS with Moses[6]. The Chinese sentences of TS, which have been segmented in Section 3.1, are used here. The 2,000 sentence pairs of the development set are used in the minimum error rate training (Och et al., 2003) for optimizing the MT system. At last, the MT system is evaluated on the 2,000 sentences pairs of the test set in BLEU score (Papineni et al., 2002). The evaluation results are shown in Table 2.

Table 2. Evaluation results of the MT systems using different Chinese word segmentation methods

MT systems using different segmentation methods	BLEU[%]
1-best of CRF segmenter	30.53
Our method (Combining multiple segmentation results)	31.15
Stanford Chinese segmenter	30.98
NLPIR Chinese segmenter	30.56

[6] http://www.statmt.org/moses/

For comparison, we also use Stanford Chinese segmenter[7] and NLPIR Chinese segmenter (ICTCLAS 2013)[8] for Chinese word segmentation of the same data and develop MT systems using the segmented data, respectively. The evaluation results on the same testing data are also shown in Table 2.

In this evaluation, we take the MT system using the result of 1-best of CRF segmenter as baseline system, whose BLEU score is 30.53%. Compared with the baseline system, the BLEU score of the MT system using the results of our method is improved by 0.62%. It proves that the proposed method not only increases the accuracy of Chinese word segmentation, but also achieves improvement in the translation quality. The BLEU scores of the MT systems using Stanford Chinese segmenter and NLPIR Chinese segmenter are 30.98% and 30.56%, respectively. So the performance of the proposed method is better than those of the two popular segmenters.

4 Conclusion

In this paper, we implement two domain-adapted Chinese word segmenters, one based on n-gram statistical feature of large Chinese raw corpus and the other one based on bilingually motivated features of parallel corpus. To augment domain adaptation, we propose a method of combining multiple Chinese segmentation results of the two segmenters based on linear model. The proposed method also provides a solution to the problem of poor performance of Chinese word segmentation in development of Chinese-English machine translation system. The experimental results on the NTCIR-10 show that both F-measure of Chinese word segmentation result and the BLEU score of the machine translation system are improved.

However, the proposed method has its limitation. The proposed method cannot be used in the application of the Chinese-English MT system for segmenting input Chinese sentences, because the corresponding English sentences are required by the bilingually motivated segmenter. In the future, we attempt to train the CRF-based segmentation model using the result of the bilingually motivated segmenter too and accordingly to achieve a word segmentation model integrated with the bilingual motivated features for application of MT system.

References

1. Zhang, M., Deng, Z., Che, W., et al.: Combining Statistical Model and Dictionary for Domain Adaption of Chinese Word Segmentation. Journal of Chinese Information Processing 26(2), 8–12 (2012)
2. Wang, Y., Kazama, J., Tsuruoka, Y., et al.: Improving Chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data. In: Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 309–317 (2011)

[7] http://nlp.stanford.edu/software/segmenter.shtml
[8] http://ictclas.nlpir.org/

3. Guo, Z., Zhang, Y., Su, C., Xu, J.: Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation. In: Zhou, M., Zhou, G., Zhao, D., Liu, Q., Zou, L. (eds.) NLPCC 2012. CCIS, vol. 333, pp. 121–131. Springer, Heidelberg (2012)
4. Ma, Y., Way, A.: Bilingually motivated domain-adapted word segmentation for statistical machine translation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 549–557. Association for Computational Linguistics (2009)
5. Xi, N., Li, B., et al.: A Chinese Word Segmentation for Statistical Machine translation. Journal of Chinese Information Processing 26(3), 54–58 (2012)
6. Ma, Y., Zhao, T.: Combining Multiple Chinese Word Segmentation Results for Statistical Machine Translation. Journal of Chinese Information Processing 1, 104–109 (2010)
7. Feng, H., Chen, K., Deng, X., et al.: Accessor variety criteria for Chinese word extraction. Computational Linguistics 30(1), 75–93 (2004)
8. Low, J.K., Ng, H.T., Guo, W.: A Maximum Entropy Approach to Chinese Word Segmentation. In: Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing (SIGHAN 2005), pp. 161–164 (2005)
9. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C++. Cambridge University Press, Cambridge (2002)
10. Xia, F.: The segmentation guidelines for the Penn Chinese Treebank (3.0). Technical report, University of Pennsylvania (2000)
11. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1, pp. 160–167. Association for Computational Linguistics (2003)
12. Papineni, K., Roukos, S., Ward, T., et al.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational linguistics, pp. 311–318 (2002)