Adapting Chinese Word Segmentation for Translation by Using a Bilingual Dictionary

Hailong Cao **Masao Utiyama** **Eiichiro Sumita**

National Institute of Information and Communications Technology
{hlcao, mutiyama, eiichiro.sumita }@nict.go.jp

Abstract

This paper proposes a method to adapt Chinese word segmentation for statistical machine translation. Two kinds of information are used to segment the Chinese sentences in the Chinese-English bilingual corpus which is the training set of the machine translation model. One is the manually segmented monolingual corpus which is widely used by general purpose segmenters. The other is the information hidden in the corresponding English sentences. In order to use the English information, rather than performing word alignment which is time consuming, we exploit a bilingual dictionary in a dynamic way. We demonstrate the usefulness of our approach on a Chinese to English translation task in a small and a large data environment.

1 Introduction

Chinese word segmentation (CWS) is a necessary step in Chinese-English statistical machine translation (SMT) and its performance has an impact on the results of SMT. The common solution in Chinese-to-English translation has been to segment the Chinese text using an off-the-shelf segmenter which is trained on a manually segmented corpus. However, the domain of the segmented corpus may not exactly match with the SMT task at hand. Consequently, the disambiguation ability of the segmenter will drop and the performance of the SMT system will be influenced.

Some work has been done to adapt CWS by using training data of SMT, i.e., bilingual parallel corpora. The segmentation ambiguity in a Chinese sentence could be resolved by referring to the corresponding English sentence in the parallel corpus. In order to use the English information, some kind of connection has to be made between the Chinese side and the English side. In most existing research, this connection is made by performing automatic alignment between Chinese characters and English words. Though alignment has been shown very effective, it is computational expensive especially on large scale corpora. Instead of using alignment information, this paper proposes a CWS method based on a bilingual dictionary. SMT experiments on both small and large corpora demonstrate the usefulness of our CWS method.

The rest of this paper is organized as follows. In section 2, we introduce our baseline CWS system and SMT system which we use in our experiments. In section 3 and 4, we describe the bilingual dictionary based CWS method and experiments on IWSLT corpus. Section 5 tests our method on NIST corpus. Related work is reviewed in section 6. Finally, conclusion is made in section 7.

2 Baseline System

2.1 CWS model

The model of our baseline segmenter is the dictionary-based bigram model (Zhang et al., 2008). We select the dictionary-based model because it is simple and effective.

The model is trained on the widely used segmentation corpus created by Peking University (PKU)[1]. The PKU corpus contains 19,056 sentences and 1,109,947 words. A dictionary which contains 55,303 unique words is extracted from the PKU corpus. A language model is built on the PKU corpus with the SRI language modeling toolkit[2]. Note that a dictionary and a language model are the only needed resources for building a dictionary-based segmenter.

[1] http://www.sighan.org/bakeoff2005/

[2] http://www.speech.sri.com/projects/srilm/

2.2 SMT model

Our SMT system is based on a fairly typical phrase-based model (Finch and Sumita, 2008). For the training of our SMT model, we use a modified training toolkit adapted from the MOSES decoder (Koehn, 2007). Our decoder can operate on the same principles as the MOSES decoder. We use a 5-gram language model trained with modified Knesser-Ney smoothing. The language model is trained on the target side of IWSLT 2008 BTEC training corpus. The translation model is created from the IWSLT 2008 BTEC training corpus. Minimum error rate training (MERT) with respect to BLEU score is used to tune the decoder's parameters, and it is performed using the standard technique of Och (2003). We use develop set 1, 2 and 4 from BTEC corpus for tuning model parameters. Develop set 3, 5 and 6 are used for testing, we call the latter test data, set3, set5 and set6 in this paper. Table 1 shows the statistics of BTEC corpus.

Data	Sentence pairs	Chinese words	English words
training set	19972	151338	161039
set1+set2+set4	1495	11369	11895
set3	506	3284	3260
set5	500	5701	6295
set6	489	2828	3142

Table 1: Statistics of BTEC corpus

The performance of SMT is evaluated by the case sensitive BLEU and NIST scores. The baseline row of Table 2 shows the SMT performance.

Test Data	Methods	BLEU	NIST
set3	Baseline	0.4623	8.2356
	Naïve	0.4753	8.0203
	Proposed	**0.4792**	**8.6037**
set5	Baseline	0.1583	4.6066
	Naïve	0.1473	4.3129
	Proposed	**0.1655**	**4.8355**
set6	Baseline	0.2564	5.2647
	Naïve	0.2559	5.2775
	Proposed	**0.2719**	**5.4943**

Table 2: Comparison of three CWS method by SMT quality

3 The Naïve Way of Using Bilingual Dictionary in CWS for SMT

In this paper, we will exploit a bilingual dictionary to adapt CWS for SMT. The bilingual dictionary that we use is a fairly large English-Chinese wordlist (LDC, 2002). It consists of a list of English words, each of which is followed by translations in Chinese separated by slashes. It has been compiled from a set of diverse resources, partly LDC-internal but mainly from the Internet. Here are some examples from it:

my	/我的/表示亲切的招呼语/吾/
wallet	/钱袋/皮夹子/小工具袋/小袋/万宝囊/旅行袋/钱包/
was	/是/
taken	/购买/
by	/在侧/经/等到...已经/因/以/依据/每.../只/在旁/向旁边/经过/过去/经由/靠/由/
a	/一个/
pickpocket	/扒手/
peridium	/包被/

A naïve way of using this bilingual dictionary is extract all the Chinese words and use them as an additional dictionary for the segmenter. We extract 142,913 unique Chinese words from it. Then all Chinese sentences in BTEC corpus are segmented by the segmenter equipped with this additional dictionary and the basic dictionary which is extracted from the PKU corpus. Figure 1 shows the pseudo code of the naive method.

```
$BD$ = Extract(PKU); # Basic Dictionary
$AD$ = Extract(bilingual dictionary); # Additional Dictionary
$Dictionary$ = Merge($BD$, $AD$);
$LM$ = SRI_TOOLIT(PKU) # Language Model
For each Chinese sentence $C$ in BTEC corpus
{
    Segment $C$ with $Dictionary$ and $LM$ ;
}
```

Figure 1: The naïve way of using bilingual dictionary in CWS for SMT

The "naïve" row of Table 2 shows the resulting SMT performance. According the BLEU measure, the additional dictionary improves over the baseline on develop set 3 but it hurts the performance on develop set 5 and 6. In order to find the reason of the inconsistent SMT performance, we manually compare the segmentation results with those that we get in the baseline experiment. We find that more Chinese words have been recognized thanks to the additional dictionary. Most of the newly recognized Chinese words are correct, but there are also some errors among them because there is much more ambiguity introduced by the additional dictionary. For example, given a Chinese sentence “我的钱包被一个扒手偷走了” whose English translation is “my wallet was taken by a pickpocket”, the baseline and the new segmentation result are “我 的 钱包 被 一 个 扒 手 偷走 了” and “我 的 钱 包被 一个 扒 手 偷走 了” respectively. It could be benefit by recognizing the word “扒手” rather than segmenting them into two individual characters. But, the word “包被” in the additional dictionary brings a new ambiguity to the character sequence “钱包被” and the model fails to resolve it due to data sparseness.

4 CWS by Using Dictionary Entries Selected Dynamically

According to the analysis in the previous section, we should limit the size of the dictionary used by the segmenter. Instead of simply extracting all Chinese words from the bilingual dictionary, we investigate to use the resource in a dynamic way. For each Chinese sentence in the training set of our SMT system, we extract dictionary entries from the bilingual dictionary by referring to the corresponding English sentence. Figure 2 shows the pseudo code of our method.

```
BD = Extract(PKU); # Basic Dictionary
LM = SRI_TOOLIT(PKU) # Language Model
For each sentence pair (C,E) in the Chinese-English parallel corpus
{
  DD = empty; # Dynamic Dictionary
  For each word e in E
  {
    If (e exist in bilingual dictionary)
    { For each word c ∈ translations of e
        {Add c into DD; }
    }
  }
  Dictionary = Merge(BD, DD);
  Segment C with Dictionary and LM ;
}
```

Figure 2: CWS by using bilingual dictionary dynamically

For example, when we segment the Chinese sentence “我的钱包被一个扒手偷走了”, we can extract a Chinese dictionary by looking up each English word of its English translation, i.e., “my”, “wallet”, “was”, “taken”, “by”, “a” and “pickpocket”, in the bilingual dictionary. Thus we get a dynamic dictionary which contains the following words:

/我的/表示亲切的招呼语/吾/钱袋/皮夹子/小工具袋/小袋/万宝囊/旅行袋/钱包/是/购买/在侧/经/等到...已经/因/以/依据/每.../只/在旁/向旁边/经过/过去/经由/靠/由/一个/扒手/

This dynamic dictionary consists of only 29 words. Clearly, it is much smaller than the dictionary used in previous section which consists of 142,913 words. The word “扒手” is included into the dynamic dictionary while the word “包被” is excluded because the word “peridium” is not in the English translation. The Chinese sentence is segmented by the segmenter equipped with the basic dictionary augmented by the dynamic dictionary. In this way, the segmentation result is “我 的 钱包 被 一个 扒手 偷走 了”. Obviously, this segmentation result is better than “我 的 钱包 被 一个 扒 手 偷走 了” or “我 的 钱 包被 一个 扒手 偷走 了” which we get in the previous two sections.

Usually, there is more than one Chinese word corresponding to one English word, but we do not perform any disambiguation and simply keep all the candidate Chinese words. In most cases, the unrelated entries such as “钱袋” and “皮夹子” do not occur in the sentence to segment, therefore little ambiguity will be introduced by them. Stemming is done if an English word is not found in the bilingual dictionary and it is limited to very simple operations such as removing –s,-ed and -ing.

By extracting a dynamic dictionary for each sentence, we segment all the Chinese sentences in the training set of our SMT model. Then we add all the automatically segmented sentences into the training corpus of our segmenter. Finally, we retrain our segmenter on the enlarged corpus and segment all the Chinese sentences in the development set and test set. The “proposed” row of Table 2 shows the resulting SMT performance. Our CWS method improves over the baseline on all the three test sets.

Data	Baseline	Naïve	Proposed
set3	182	210	189
set5	398	442	409
set6	156	175	170

Table 3: The number of OOV word

Table 3 shows the number of out of vocabulary (OOV) word in three experiments. Clearly, the naive method generates many more OOV words than the other two methods.

In a dynamic way, the bilingual dictionary can improve the CWS for SMT. However, there is much potential room for further improvement. For example, the bilingual dictionary has indicated that there would be a word "我的" in the mentioned sentence, but the segmentation result is "我 的" because the bigram "我 的" occurs more than 100 times in the PKU corpus. Using the bilingual dictionary to adapt the word granularity of manually segmented corpus is a direction of our future work.

5 Experiment on Large Scale Data

Our method almost requires no additional computation cost and can be easily applied to large scale data. So we further evaluate the validity of our method on the NIST machine translation task (NIST, 2008). As shown in Table 4, the training corpus contains more than four million sentence pairs.

We perform the experiments described in section 2 and 4 again on the NIST 2008 data. All experimental settings are the same with that on the IWSLT corpus. The BLEU score we get in the baseline experiment is 0.1969. Our CWS method based on bilingual dictionary improves the SMT performance to 0.2020 BLEU score.

Data	Sentence pairs	Chinese words	English words
Training set	4410100	76478284	76576933
Develop set	1664	40859	46387
Test set	1357	34569	42444

Table 4: Statistics of corpora in NIST

6 Related Work

Recently, much work has been done to optimize word segmentation for SMT. Xu et al. (2005) take different segmentation alternatives instead of a single segmentation into account and integrate the segmentation process with the search for the best translation. The segmentation decision is only taken during the generation of the translation. Xu et al. (2006) propose an integration algorithm of English-Chinese word segmentation and alignment. In this work, segmentation and alignment work synchronously. Ma et al. (2007) introduce a method to pack words for word alignment. They simplify the task of word alignment by packing consecutive words together if they correspond to a single word in the opposite language. Chang et al. (2008) investigate what segmentation properties can improve SMT performance and propose an algorithm to directly optimize segmentation granularity for translation quality. Ma et al. (2009) propose a CWS method for SMT based on statistical word alignment.

The difference between the above work and ours is that we use a bilingual dictionary instead of performing word alignment automatically. Our method is applicable when a suitable bilingual dictionary is available.

7 Conclusion and Future Work

When we perform CWS for Chinese-English SMT, the information in the English side is quite useful to resolve CWS ambiguity. In this paper, a very simple method is proposed to get better CWS for SMT by making use of the information in the English side and a bilingual dictionary in a dynamic way. We demonstrate the usefulness of our method on both small corpora and large scale corpora.

In this paper, we focus on Chinese-English SMT, it should be noted that our method is language independent. In the future, we will test our method on other language pairs such as Japanese-English. Furthermore, whether a bilingual dictionary that is automatically generated, which should be large but has poor quality is useful or not, is a question should be investigated.

References

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224-232, Columbus, OH.

Andrew Finch and Eiichiro Sumita. 2008. Dynamic model interpolation for statistical machine translation. In *proceedings of the Third Workshop on Statistical Machine Translation*, pages 208-215.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation**.** In *proceedings of ACL demo and poster sessions*, Prague, Czech Republic, pages 177-180.

LDC. 2002. *English-to-Chinese Wordlist* (version 2.). http://www.ldc.upenn.edu/Projects/Chinese/.

Yanjun Ma and Andy Way. 2009. Bilingually Motivated Domain-Adapted Word Segmentation for Statistical Machine Translation. In *Proceedings of the EACL*, Athens, Greece.

Yanjun Ma, Nicolas Stroppa and Andy Way. 2007. Bootstrapping Word Alignment via Word Packing. In *Proceedings of ACL*, pages 304-311.

Franz Och. Minimum error rate training in statistical machine translation. 2003. In *Proceedings of ACL*, pages 160-167.

Jia Xu, Evgeny Matusov, Richard Zens and Hermann Ney. 2005. Integrated Chinese Word Segmentation in Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 141-147.

Zhiming Xu, Chunyu Kit and Jonathan J. Webster. 2006. Integration algorithm of English-Chinese word segmentation and alignment. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pages 5105-4110.

Ruiqiang Zhang, Keiji Yasuda and Eiichiro Sumita. 2008. Improved statistical machine translation by multiple Chinese word segmentation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 216-223.