Using Parallel Corpora to Automatically Generate Training Data for Chinese Segmenters in NTCIR PatentMT Tasks

Jui-Ping Wang Chao-Lin Liu
Department of Computer Science, National Chengchi University, Taipei, Taiwan
{99753016, chaolin}@nccu.edu.tw

Abstract
Chinese texts do not contain spaces as word separators like English and many alphabetic languages. To use Moses to train translation models, we must segment Chinese texts into sequences of Chinese words. Increasingly more software tools for Chinese segmentation are populated on the Internet in recent years. However, some of these tools were trained with general texts, so might not handle domain-specific terms in patent documents very well. Some machine-learning based tools require us to provide segmented Chinese to train segmentation models. In both cases, providing segmented Chinese texts to refine a pre-trained model or to create a new model for segmentation is an important basis for successful Chinese-English machine translation systems. Ideally, high-quality segmented texts should be created and verified by domain experts, but doing so would be quite costly. We explored an approach to algorithmically generate segmented texts with parallel texts and lexical resources. Our scores in NTCIR-10 PatentMT indeed improved from our scores in NTCIR-9 PatentMT with the new approach.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language generation, language models, machine translation*; I.2.6 [**Artificial Intelligence**] Learning – *knowledge acquisition, parameter learning*; H.2.8 [**Database Management**]: Database Applications – *data mining*.

General Terms
Algorithms, Experimentation, Languages

Keywords
Chinese-English Patent Machine Translation, Chinese Near Synonyms, Chinese Segmentation, Machine Learning

Team Name
[MIG]

Subtasks/Languages
[Intrinsic Evaluation], [Adequacy], [Chronological Evaluation], [Multilingual Evaluation] / [Chinese-English]

External Resources Used
Moses, Stanford Chinese segmenter, LingPipe Chinese segmenter, Chinese lexicons, English-Chinese lexicons

1. Introduction
A common architecture for Chinese-English machine translation systems considers three models: segmentation model, translation model, and language model. In the segmentation model, we separate the source Chinese strings into individual words. In the translation model, we convert the Chinese words into possible English translations. Finally, we determine the best orders for the English words. As a major step for processing Chinese texts, the quality of Chinese segmentation must influence the quality of Chinese-English machine translation [9].While the most intuitive way to organize these steps in the presented order, it is possible for one to take an iterative approach to optimize the overall performance.

In our work for NTCIR-9 PatentMT task [6] , we relied on Moses[1] to handle most of the computation for statistical machine translation (SMT). We employed two tools for Chinese segmentation. The first was the Chinese Segmenter included in the natural language processing tools of the Stanford University[2] [2], and the second was offered by the LingPipe System[3]. We refer to these two tools as SCS (standing for Stanford Chinese Segmenter) and LPS (standing for LingPipe Segmenter), respectively.

Both SCS and LPS allow us to train segmentation models with our training data. For the NTCIR-9 PatentMT task, we relied on a proprietary procedure offered by the WebGenie Corporation[4] to segment the Chinese sentences. For the NTCIR-10 PatentMT task, we explored methods for automatic generation of data for training **SCS** and **LPS**.

We used the default settings of Moses when we did the translation, and changed only how we segmented the Chinese texts which were used to train the models in Moses. Hence, we have the opportunity to compare the influences of different ways to segment Chinese texts on the final tests in two versions of the PatentMT tasks [4].

The official results of the final tests show that the new approach led to better BLEU scores in NTCIR-10. We motivate and present our methods for creating the training data in Section 2, explain the language resources used in the methods in Section 3, report results of internal evaluation of our models in Section 4, and present the final results in Section 5 before we make concluding remarks in Section 6.

2. Training Data Generation

2.1 Motivation
The PatentMT tasks provided parallel corpora to participants to train their systems for machine translation. Parallel corpora are useful in multiple ways for natural language processing [1]. For

[1] http://www.statmt.org/moses/
[2] http://nlp.stanford.edu/software/segmenter.shtml
[3] Alias-i. 2008. LingPipe 4.1.0. http://alias-i.com/lingpipe
[4] http://www.webgenie.com.tw/

instance, using good techniques for word alignment, one may create a bilingual lexicon with parallel corpus [8].

For our interests in both NTCIR-9 and NTCIR-10, we compared the effects of different ways to segment Chinese texts to obtain training data for Moses.

The ideal solution for training a high-quality Chinese segmenter is to obtain high-quality training data in the first place. If one affords to have human experts segment Chinese texts and verify the results, one may achieve impressive segmentation results with good machine learning algorithms as those winners in SIGHAN BackOff[5].

In reality, it is not easy to obtain a large amount of high-quality segmented Chinese text. Moreover, for applications that are domain dependent, application builders must face the problems of scarcity and availability of training data.

2.2 The Main Procedure

We illustrate the main idea for automatic creation of training data for Chinese segmenters, using the parallel corpus of the PatentMT task in the flow in Figure 1.

For each of the 1 million Chinese-English sentence pairs, we first processed the Chinese sentences. We collected Chinese lexicons that were useful for the problem domains for which we would do machine translation. We also needed Chinese lexicons for everyday terms. With these lexicons, we could produce all possible segmentations for each Chinese sentence [7].

The second rectangle in Figure 2 shows two possible segmentations of the Chinese sentence of an English-Chinese sentence pair. We use slashes to divide Chinese strings into sequences of words. In this example, both "即便" and "便是" were known Chinese words in the Chinese lexicons. We had no good reasons to prefer an alternative when we examined only the Chinese sentence, so

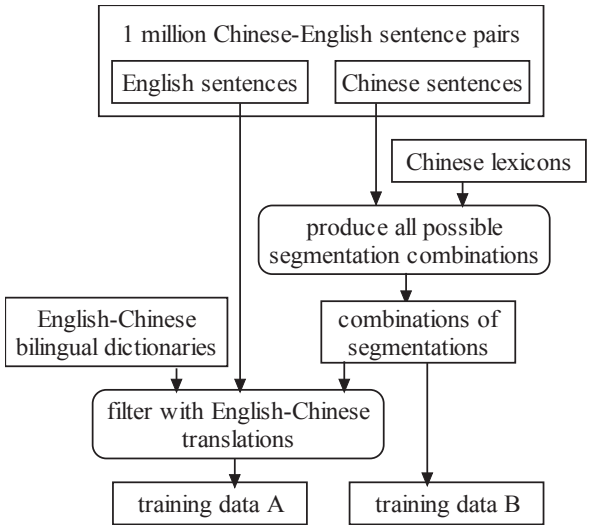

Figure 1. Creating training data with bilingual resources

we kept both of them at least for now.

The left side of Figure 1 shows how we may employ English information to filter Chinese segmentations. We could look up the English-Chinese dictionaries to find possible Chinese translations of the words in the English sentence. This is shown at steps 1 and 2 in Figure 2.

Since the sentence pairs in NTCIR-10 PatentMT corpus are translated pairs, we can use the Chinese translations of the English words to constrain the Chinese segmentations.

Step 3 in Figure 2 presents an example. Because "即便" and "便是" are both possible Chinese words, we kept them in a previous step. Now because we have "even" in the English sentence and one of its Chinese translations is "即便", we would tend to adopt

Input: an English-Chinese sentence pair
English sentence: Even those projects can induce earthquakes, although most are small. Chinese sentence: 即便是這類工程都可能引發地震，不過大半規模不大。

multiple Chinese segmentations
Chinese segmentation 1：即便/是/這/類/工程/都/可能/引發/地震/，/不過/大半/規模/不大/。 Chinese segmentation 2：即/便是/這/類/工程/都/可能/引發/地震/，/不過/大半/規模/不大/。

Step 1：Lemmatize English words in the English sentence
even、those、project、 can、 induce、 earthquake、 although 、most、 be、 small

Step 2：Find the Chinese translations of English words, including Chinese near synonyms		
even：縱使, 縱然, 即便…	those：那些, 那	project：計劃, 方案, 事業…
can：可能, 會, 可以…	induce：勸誘, 促使, 導致…	earthquake：地震, 大動盪, 天搖地動…
although：雖然, 儘管, 即使…	most：至多, 頂多, 最…	be：位於, 身處, 是…
small：小的, 少的, 小型的…		

Step 3：Apply the English-Chinese word translation pairs to filter Chinese segmentations
Chinese segmentation 1：<even/即便>即便/是/這/類/工程/都/可能/引發/地震/，/不過/大半/規模/不大/。 Chinese segmentation 2：~~即/便是/這/類/工程/都/可能/引發/地震/，/不過/大半/規模/不大/。~~

Output
即便/是/這/類/工程/都/可能/引發/地震/，/不過/大半/規模/不大/。

Figure 2. An illustration of using English-Chinese word translations to filter Chinese segmentations

[5] http://www.sighan.org/bakeoff2006/

"即便" rather than "便是" in the segmentation choice. Hence we can rule out the second Chinese segmentation in Figure 2.

2.3 Competing Segmentations

The procedure that we illustrated with Figures 1 and 2 was designed to handle the ***competing segmentations*** in which two or more segmentation alternatives compete for a common substring. The most simple example of competing segmentations is that we may segment the string ABC into AB/C or A/BC, where A, B, and C represents a Chinese character.

"即便是" in Figure 2 is an example, and "一旦有機會" is another example. We can segment the latter in two different ways: "一旦/有/機會" and "一旦/有機/會". The situation is similar to the associativity problem in the design of programming languages – whether the character "機" should be associative with its left or right context. If we have information about its English translation, we would have a better, if not precise, idea about how we should segment this string. If the English translation has "chance", then the former segmentation is more possible. If the English translation has "organic", then the latter is more reasonable.

2.4 Near Synonyms in Chinese

If our bilingual dictionary does not translate an English word exactly the same way as the patent documents did, the procedure we depicted in Figure 1 would not work. Hence, the aforementioned procedure demands an important step of preparation. How did we know that the word translations in an English-Chinese dictionary are also used in the English-Chinese sentence pairs? Would it be possible that the patent documents employed conceptually equivalent ways to translate the same concepts? If an English word has multiple ways to be translated into Chinese, how can we make sure that an English-Chinese dictionary that we use indeed includes all of the possible translations? This is something that is hard to be guaranteed.

As a result, we tried to collect as many bilingual dictionaries as possible. Having more dictionaries offer ourselves more comprehensive lists of translations of English words.

In addition, we relied on external sources where Chinese near-synonyms were provided, e.g. the web site of "一詞泛讀"[6] maintained by the Academia Sinica. Moreover, we employed the E-HowNet to compute near synonyms with our own methods [3].

We used the near synonyms as if they were acceptable translations under any circumstances. Namely, in Figure 1, we would use the original Chinese words, the Chinese translations in the bilingual dictionaries, and their near synonyms to match the Chinese strings in the Chinese sentences.

This, of course, is not always right, particularly for ambiguous words. If "bank" appears in the English part of a pair of English-Chinese sentence, we may consider "依賴" in the Chinese part as the translation of "bank" without considering the context of the Chinese sentence. It is "rely" in "… we relied on the loan from the First Bank to buy this house …" to be translated to "依賴" rather than "bank."

To avoid the above problem completely, precise word alignment is in need, which is still beyond our capability. We hope problems of this type are not very common in practical patent documents, and our current approach takes the risk to ignore such potential problems.

Table 1. Chinese lexicons

Category	Name	Types
Ordinary	MOEDICT	157704
Ordinary	CYCD	13947
Ordinary	GJHYDCD	54467
Professional	CTTD	804053
Professional	WCE	648612

Table 2. English-Chinese lexicons

Category	Name	Type (English)	Translation (Chinese)
Ordinary	CEC	99805	3729292
Ordinary	Lazyworm	121525	323766
Professional	ECDTT	586075	804053

2.5 Out-Of-Vocabulary Words

The existence of Out-Of-Vocabulary (OOV) words is a very challenging issue for Chinese segmentation. Sources of OOV words include very basic words that ordinary Chinese dictionaries do not bother to include. OOV words may also come from new words which is particularly possible for Chinese patent documents.

For the former class of OOV words, one may embrace statistical methods along with elementary methods for word alignment to identify the basic words [5]. Words of this category appear frequently in Chinese texts, so statistical methods provide good opportunities to identify them.

It is relatively difficult to identify OOV words that are both new and infrequent. This is because of the fact that patent documents often introduce innovative technologies and new products.

Although we can handle OOV words partially, we did not apply this capability when we prepare our submissions for NTCIR-10.

3. Lexicons

The quality of data generated by the procedure sketched in Figure 1 relies on the quality of the Chinese lexicons and the English-Chinese lexicons.

Table 1 lists the Chinese lexicons used in our work. We considered two categories of lexicons: ordinary and professional. The first ordinary lexicon was obtained from the Ministry of Education (**MOEDICT**) of Taiwan, and MOEDICT has about 157 thousand different words. ("Types" in Table 1 indicate "different words" in natural language processing.) The second lexicon is for Chinese idioms – *Cheng Yu Ci Dian* (**CYCD**, *成語詞典*). The third is an advanced Chinese lexicon (*Gao Ji Han Yu Da Ci Dian*, **GJHYDCD**, *高級漢語大詞典*). The fourth came from the Chinese part of an English-Chinese dictionary for technical terms, which we explain shortly. We use **CTTD** (Chinese Technical Term Dictionary) to refer to this dictionary. The last one contains translated (Chinese) names of world-class elites (*世界人名翻譯大辭典* [7]), and we refer to this dictionary as **WCE**.

Table 2 lists our English-Chinese lexicons. We collected the English-Chinese entries from both Dr.Eye and the Concise Oxford

[6] http://elearning.ling.sinica.edu.tw/c_help.html

[7] http://zh.wikipedia.org/zh-hant/世界人名翻譯大辭典

Table 3. Four translation models

Segmenter	Training Data	Translation Model
LPS	B	TM1
LPS	A	TM2
SCS	B	TM3
SCS	A	TM4

Table 4. Experimental results: NTCIR-9 tuning data

Translating Model	NIST	BLEU
TM1	7.5800	0.2861
TM2	7.4608	0.2738
TM3	7.4067	0.2715
TM4	7.5679	0.2843

English Dictionary, and expanded the Chinese translations with their near synonyms as we discussed in Section 2.4. We refer to this combined lexicon as **CEC**. We obtained the Lazyworm dictionary[8] from the Internet. We acquired an English-Chinese dictionary of technical terms (**ECDTT**) from the National Academy for Educational Research of Taiwan. ECDTT contains technical terms of 138 different fields. We extracted the Chinese part of ECDTT to form CTTD in Table 1.

4. Internal Evaluation

We ran the procedure shown in Figure 1 with the 1 million English-Chinese sentence pair distributed for NTCIR-10 PatentMT task. As the figure suggests, we obtain two types of training data for the Chinese segmenters. Training data B contains all of the possible Chinese segmentations. We would remove less appropriate Chinese segmentation from training data B, using the steps illustrated in Figure 2, to obtain training data A.

As we mentioned in Introduction, we employed SCS (Stanford Chinese Segmenter) and LPS (LingPipe Segmenter) in our experiments. Table 3 shows the four possible combinations based on which we built translation models with Moses for NTCIR-10 PatentMT task.

For example, we used training data B and the Chinese sentences of the 1 million English-Chinese sentence pairs to train a segmentation model with LPS. We then applied the segmentation model to segment the Chinese sentences in the 1 million English-Chinese sentence pairs. Finally, we used Moses to process the segmented Chinese sentences and the original English sentences to produce the translation model TM1.

Since the training data for NTCIR-10 and NTCIR-9 PatentMT tasks are the same, we tested the translation models with the test data for the NTCIR-9 PatentMT task, and observed the results in Table 4. These scores are better than those that we achieved when we prepared for the submissions for the NTCIR-9 PatentMT task [6].

It was encouraging to achieve better BLEU scores than we did last year with the simpler method for training Chinese segmenters,

At the same, we observed conflicting results in Table 4. Using the English-Chinese parallel corpus to filter Chinese segmentations led to better performance of SCS (TM3 vs. TM4). However, filtering Chinese segmentation did not help to improve the performance of LPS (TM1 vs. TM2).

[8] http://abloz.com/huzheng/stardict-dic/zh_TW/

Table 5. Final results

ID	Translation Model	BLEU
MIG-ze-int-1	TM1	0.3018
MIG-ze-int-2	TM4	0.3017
MIG-ze-int-3	TM2	0.3012
MIG-ze-int-4	TM3	0.2866
MIG-ze-chr-1	TM1	0.2861
MIG-ze-mul-1	TM1	0.1812

5. Submissions and Final Results

We ranked the translation models based on the BLEU scores that they achieved in Table 4. Based on the ranked result, we used TM1, TM4, TM2, and TM3 to translate the test data for NTCIR-10 PatentMT task, and the results were encoded, respectively, by MIG-ze-int-1, MIG-ze-int-2, MIG-ze-int-3, and MIG-ze-int-4 for the Intrinsic Evaluation.

We used TM1 to produce the submissions for Patent Examination Evaluation, Chronological Evaluation, and Multilingual Evaluation because only one submission was expected for these tasks. Our submissions were encoded, respectively, by MIG-1, MIG-ze-chr-1, and MIG-ze-mul-1 in the announcement for the formal run results.

Since our current focus was not to build a complete Patent MT system, our BLEU scores, listed in Table 5, have a big room to improve, when compared with the best performing systems in NTCIR-10.

For the Patent Examination Evaluation, we achieved 3.05 in translation adequacy on a scale of 1 to 5 [4].

For the Chronological Evaluation, the BLEU scores listed in Table 4 and Table 5 indicate that we achieved better scores in NTCIR-10. What is significant to us was that the improvement of our scores was achieved by a conceptually simpler method to train and obtain our Chinese segmenter, as we explained in Section 2.2

For the Multilingual Evaluation, our system accomplished a BLEU score, MIG-ze-mul-1 in Table 5, that was just good enough to beat the baseline systems.

6. Concluding Remarks

We took advantage of the NTCIR PatentMT task as an external evaluation of our work on Chinese segmentation. In NTCIR-9, we relied on a proprietary mechanism of Webgenie for Chinese segmentation. In NTCIR-10, we explored the effectiveness of a simple way to generate training data for publically available tools for Chinese segmentation, the Stanford Chinese Segmenter and the LingPipe Segmenter. Data created by our procedure were used to train the segmenters, and data produced by the trained segmenters were then used to create translation models by Moses. New translation models achieved higher BLEU scores than the translation models that we created for NTCIR-9, suggesting that the proposed method not only alleviates the cost problem to collect data for training Chinese segmenters but also offers chances to achieve better quality in machine translation.

Acknowledgements

This work was supported in part by the grants NSC-100-2221-E-004-014- and NSC-101-2221-E-004-018- from the National Science Council, Taiwan.

References

[1] B. Chang. Chinese-English parallel corpus construction and its application, *Proceedings of the Eighteenth Pacific Asia Conference on Language, Information, and Computation*, 283–290, 2004.

[2] P.-C. Chang, D. Jurafsky, and C. D. Manning. Optimizing Chinese word segmentation for machine translation performance, *Proceedings of the Third Workshop on Machine Translation*, 2008.

[3] Y.-H. Chuang, J.-P. Wang, C.-C. Tsai, and C.-L. Liu. Collocational influences on the Chinese translation of non-technical English verbs and their objects in technical documents, *Proceedings of the Twenty Third Conference on Computational Linguistics and Speech Processing*, 94–108, 2011.

[4] I. Goto, K. P. Chow, B. Lu, E. Sumita, and B. K. Tsou. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop, *Proceedings of NTCIR-10*, 2013.

[5] C.-S. Huang, Y.-C. Chang, C.-L. Liu, and Y.-H. Tseng. Using co-occurrence information to improve Chinese-English word alignment in translation test items for high school students, *Proceedings of the Twenty Second Conference on Computational Linguistics and Speech Processing*, 128–142, 2010.

[6] Y.-H. Tseng, C.-L. Liu, C.-C. Tsai, J.-P. Wang, Y.-H. Chuang, and J. Jeng. Statistical approaches to patent translation - Experiments with various settings of training data, *Proceedings of the NTCIR-9 - PatentMT*, 661–665, 2011.

[7] J.-P. Wang and C.-L. Liu. Applications of parallel corpora for Chinese segmentation, In *Proceedings of the Twenty Third Conference on Computational Linguistics and Speech Processing*, 341–355, 2012.

[8] D. Wu and X. Xia. Learning an English-Chinese lexicon from a parallel corpus, *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 206–213, 1994.

[9] R. Zhang, K. Yasuda, and E. Sumita. Improved statistical machine translation by multiple Chinese word segmentation, *Proceedings of the Third Workshop on Statistical Machine Translation*, 216–223, 2008.