Do We Need Chinese Word Segmentation for Statistical Machine Translation?

Jia Xu and **Richard Zens** and **Hermann Ney**
Chair of Computer Science VI
Computer Science Department
RWTH Aachen University, Germany
{xujia,zens,ney}@cs.rwth-aachen.de

Abstract

In Chinese texts, words are not separated by white spaces. This is problematic for many natural language processing tasks. The standard approach is to segment the Chinese character sequence into words. Here, we investigate Chinese word segmentation for statistical machine translation. We pursue two goals: the first one is the maximization of the final translation quality; the second is the minimization of the manual effort for building a translation system.

The commonly used method for getting the word boundaries is based on a word segmentation tool and a predefined monolingual dictionary. To avoid the dependence of the translation system on an external dictionary, we have developed a system that learns a domain-specific dictionary from the parallel training corpus. This method produces results that are comparable with the predefined dictionary.

Further more, our translation system is able to work without word segmentation with only a minor loss in translation quality.

1 Introduction

In Chinese texts, words composed of single or multiple characters, are not separated by white spaces, which is different from most of the western languages. This is problematic for many natural language processing tasks. Therefore, the usual method is to segment a Chinese character sequence into Chinese "words".

Many investigations have been performed concerning Chinese word segmentation. For example, (Palmer, 1997) developed a Chinese word segmenter using a manually segmented corpus. The segmentation rules were learned automatically from this corpus. (Sproat and Shih, 1990) and (Sun et al., 1998) used a method that does not rely on a dictionary or a manually segmented corpus. The characters of the unsegmented Chinese text are grouped into pairs with the highest value of mutual information. This mutual information can be learned from an unsegmented Chinese corpus.

We will present a new method for segmenting the Chinese text without using a manually segmented corpus or a predefined dictionary. In statistical machine translation, we have a bilingual corpus available, which is used to obtain a segmentation of the Chinese text in the following way. First, we train the statistical translation models with the unsegmented bilingual corpus. As a result, we obtain a mapping of Chinese characters to the corresponding English words for each sentence pair. By using this mapping, we can extract a dictionary automatically. With this self-learned dictionary, we use a segmentation tool to obtain a segmented Chinese text. Finally, we retrain our translation system with the segmented corpus.

Additionally, we have performed experiments without explicit word segmentation. In this case, each Chinese character is interpreted as one "word". Based on word groups, our machine translation system is able to work without a word segmentation, while having only a minor translation quality relative loss of less than 5%.

2 Review of the Baseline System for Statistical Machine Translation

2.1 Principle

In statistical machine translation, we are given a source language ('French') sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language ('English') sentence $e_1^I = e_1 \dots e_i \dots e_I$. Among all possible target language sentences, we will choose the sentence

with the highest probability:

$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I|f_1^J) \right\} \quad (1)$$

$$= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J|e_1^I) \right\} \quad (2)$$

The decomposition into two knowledge sources in Equation 2 is known as the source-channel approach to statistical machine translation (Brown et al., 1990). It allows an independent modeling of target language model $Pr(e_1^I)$ and translation model $Pr(f_1^J|e_1^I)$[1]. The target language model describes the well-formedness of the target language sentence. The translation model links the source language sentence to the target language sentence. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. We have to maximize over all possible target language sentences.

The resulting architecture for the statistical machine translation approach is shown in Figure 1 with the translation model further decomposed into lexicon and alignment model.

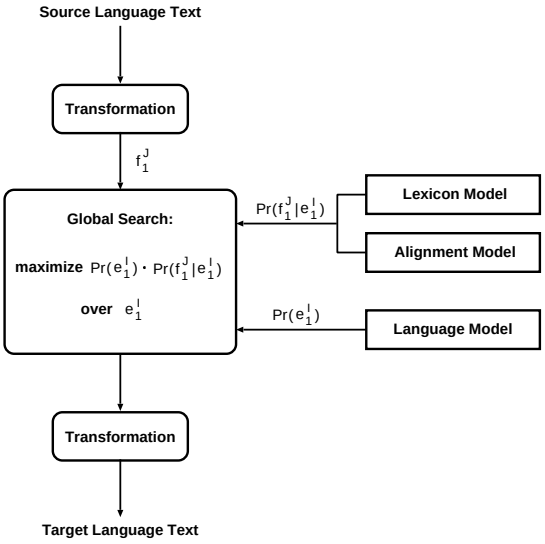

Figure 1: Architecture of the translation approach based on Bayes decision rule.

2.2 Alignment Models

The alignment model $Pr(f_1^J, a_1^J|e_1^I)$ introduces a 'hidden' alignment $\mathbf{a} = a_1^J$, which describes a mapping from a source position j to a target position a_j. The relationship between the translation model and the alignment model is given by:

$$Pr(f_1^J|e_1^I) = \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \quad (3)$$

In this paper, we use the models IBM-1, IBM-4 from (Brown et al., 1993) and the Hidden-Markov alignment model (HMM) from (Vogel et al., 1996). All these models provide different decompositions of the probability $Pr(f_1^J, a_1^J|e_1^I)$. A detailed description of these models can be found in (Och and Ney, 2003).

A *Viterbi alignment* $\hat{a}_1^J$ of a specific model is an alignment for which the following equation holds:

$$\hat{a}_1^J = \underset{a_1^J}{\operatorname{argmax}} \; Pr(f_1^J, a_1^J|e_1^I). \quad (4)$$

The alignment models are trained on a bilingual corpus using GIZA++(Och et al., 1999; Och and Ney, 2003). The training is done iteratively in succession on the same data, where the final parameter estimates of a simpler model serve as starting point for a more complex model. The result of the training procedure is the Viterbi alignment of the final training iteration for the whole training corpus.

2.3 Alignment Template Approach

In the translation approach from Section 2.1, one disadvantage is that the contextual information is only taken into account by the language model. The single-word based lexicon model does not consider the surrounding words. One way to incorporate the context into the translation model is to learn translations for whole word groups instead of single words. The key elements of this translation approach (Och et al., 1999) are the *alignment templates*. These are pairs of source and target language phrases with an alignment within the phrases.

The alignment templates are extracted from the bilingual training corpus. The extraction algorithm (Och et al., 1999) uses the word alignment information obtained from the models in Section 2.2. Figure 2 shows an example of a word aligned sentence pair. The word alignment is represented with the black boxes. The figure also includes some of the possible alignment templates, represented as the larger, unfilled rectangles. Note that the extraction algo-

[1]The notational convention will be as follows: we use the symbol $Pr(\cdot)$ to denote general probability distributions with (nearly) no specific assumptions. In contrast, for model-based probability distributions, we use the generic symbol $p(\cdot)$.

rithm would extract many more alignment templates from this sentence pair. In this example, the system input was the sequence of Chinese characters *without* any word segmentation. As can be seen, a translation approach that is based on phrases circumvents the problem of word segmentation to a certain degree. This method will be referred to as *"translation with no segmentation"* (see Section 5.2).

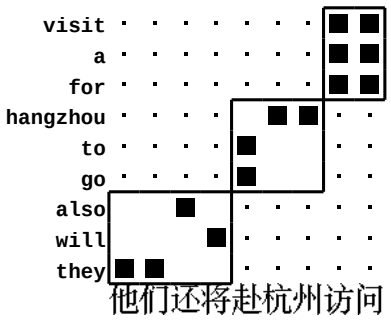

Figure 2: Example of a word aligned sentence pair and some possible alignment templates.

In the Chinese–English DARPA TIDES evaluations in June 2002 and May 2003, carried out by NIST (NIST, 2003), the alignment template approach performed very well and was ranked among the best translation systems.

Further details on the alignment template approach are described in (Och et al., 1999; Och and Ney, 2002).

3 Task and Corpus Statistics

In Section 5.3, we will present results for a Chinese–English translation task. The domain of this task is news articles. As bilingual training data, we use a corpus composed of the English translations of a Chinese Treebank. This corpus is provided by the Linguistic Data Consortium (LDC), catalog number LDC2002E17. In addition, we use a bilingual dictionary with 10K Chinese word entries provided by Stephan Vogel (LDC, 2003b).

Table 1 shows the corpus statistics of this task. We have calculated both the number of words and the number of characters in the corpus. In average, a Chinese word is composed of 1.49 characters. For each of the two languages, there is a set of 20 special characters, such as digits, punctuation marks and symbols like "()%$..."

The training corpus will be used to train a word alignment and then extract the alignment templates and the word-based lexicon. The resulting translation system will be evaluated on the test corpus.

Table 1: Statistics of training and test corpus. For each of the two languages, there is a set of 20 special characters, such as digits, punctuation marks and symbols like "()%$..."

		Chinese	English
Train	Sentences	4 172	
	Characters	172 874	832 760
	Words	116 090	145 422
	Char. Vocab.	3 419 + 20	26 + 20
	Word Vocab.	9 391	9 505
Test	Sentences	993	
	Characters	42 100	167 101
	Words	28 247	26 225

4 Segmentation Methods

4.1 Conventional Method

The commonly used segmentation method is based on a segmentation tool and a monolingual Chinese dictionary. Typically, this dictionary has been produced beforehand and is independent of the Chinese text to be segmented. The dictionary contains Chinese words and their frequencies. This information is used by the segmentation tool to find the word boundaries. In the LDC method (see Section 5.2) we have used the dictionary and segmenter provided by the LDC. More details can be found on the LDC web pages (LDC, 2003a). This segmenter is based on two ideas: it prefers long words over short words and it prefers high frequency words over low frequency words.

4.2 Dictionary Learning from Alignments

In this section, we will describe our method of learning a dictionary from a bilingual corpus.

As mentioned before, the bilingual training corpus listed in Section 3 is the only input to the system. We firstly divide every Chinese characters in the corpus by white spaces, then train the statistical translation models with this unsegmented Chinese text and its English translation, details of the training method are described in Section 2.2.

To extract Chinese words instead of phrases as in Figure 2, we configure the training pa-

rameters in GIZA++, the alignment is then restricted to a multi-source-single-target relationship, i.e. one or more Chinese characters are translated to one English word.

The result of this training procedure is an alignment for each sentence pair. Such an alignment is represented as a binary matrix with $J \cdot I$ elements.

An example is shown in Figure 3. The unsegmented Chinese training sentence is plotted along the horizontal axes and the corresponding English sentence along the vertical axes. The black boxes show the Viterbi alignment for this sentence pair. Here, for example the first two Chinese characters are aligned to "industry", the next four characters are aligned to "restructuring".

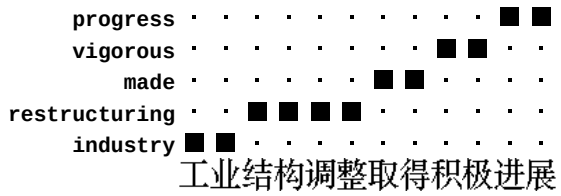

Figure 3: Example of an alignment without word segmentation.

The central idea of our dictionary learning method is: *a contiguous sequence of Chinese characters constitute a Chinese word, if they are aligned to the same English word.* Using this idea and the bilingual corpus, we can automatically generate a Chinese word dictionary. Table 2 shows the Chinese words that are extracted from the alignment in Figure 3.

Table 2: Word entries in Chinese dictionary learned from the alignment in Figure 3.

Nr.	Entry
1	工业
2	结构调整
3	取得
4	积极
5	进展

We extract Chinese words from all sentence pairs in the training corpus. Therefore, it is straightforward to collect word frequency statistics that are needed for the segmentation tool. Once, we have generated the dictionary, we can produce a segmented Chinese corpus using the method described in Section 4.1. Then, we retrain the translation system using the segmented Chinese text.

4.3 Word Length Statistics

In this section, we present statistics of the word lengths in the LDC dictionary as well as in the self-learned dictionary extracted from the alignment.

Table 3 shows the statistics of the word lengths in the LDC dictionary as well as in the learned dictionary. For example, there are 2 368 words consisting of a single character in learned dictionary and 2 511 words in the LDC dictionary. These single character words represent 16.9% of the total number of entries in the learned dictionary and 18.6% in the LDC dictionary.

We see that in the LDC dictionary more than 65% of the words consist of two characters and about 30% of the words consist of a single character or three or four characters. Longer words with more than four characters constitute less than 1% of the dictionary. In the learned dictionary, there are many more long words, about 15%. A subjective analysis showed that many of these entries are either named entities or idiomatic expressions. Often, these idiomatic expressions should be segmented into shorter words. Therefore, we will investigate methods to overcome this problem in the future. Some suggestions will be discussed in Section 6.

Table 3: Statistics of word lengths in the LDC dictionary and in the learned dictionary.

word length	LDC dictionary		learned dictionary	
	frequency	[%]	frequency	[%]
1	2 334	18.6	2 368	16.9
2	8 149	65.1	5 486	39.2
3	1 188	9.5	1 899	13.6
4	759	6.1	2 084	14.9
5	70	0.6	791	5.7
6	20	0.2	617	4.4
7	6	0.0	327	2.3
≥8	11	0.0	424	3.0
total	12 527	100	13 996	100

5 Translation Experiments

5.1 Evaluation Criteria

So far, in machine translation research, a single generally accepted criterion for the evaluation of the experimental results does not exist. We have used three automatic criteria. For the test corpus, we have four references available. Hence, we compute all the following criteria with respect to multiple references.

- WER (word error rate):
  The WER is computed as the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the reference sentence.
- PER (position-independent word error rate):
  A shortcoming of the WER is that it requires a perfect word order. The word order of an acceptable sentence can be different from that of the target sentence, so that the WER measure alone could be misleading. The PER compares the words in the two sentences ignoring the word order.
- BLEU score:
  This score measures the precision of unigrams, bigrams, trigrams and fourgrams with respect to a reference translation with a penalty for too short sentences (Papineni et al., 2001). The BLEU score measures accuracy, i.e. large BLEU scores are better.

5.2 Summary: Three Translation Methods

In the experiments, we compare the following three translation methods:

- Translation with **no segmentation**: Each Chinese character is interpreted as a single word.
- Translation with **learned segmentation**: It uses the self-learned dictionary.
- Translation with **LDC segmentation**: The predefined LDC dictionary is used.

The core contribution of this paper is the method we called *"translation with learned segmentation"*, which consists of three steps:

- The input is a sequence of Chinese characters without segmentation. After the training using GIZA++, we extract a monolingual Chinese dictionary from the alignment. This is discussed in Section 4.2, and an example is given in Figure 3 and Table 2.
- Using this learned dictionary, we segment the sequence of Chinese characters into words. In other words, the LDC method is used, but the LDC dictionary is replaced by the learned dictionary (see Section 4.1).
- Based on this word segmentation, we perform another training using GIZA++. Then, after training the models IBM1, HMM and IBM4, we extract bilingual word groups, which are referred as alignment templates.

5.3 Evaluation Results

The evaluation is performed on the LDC corpus described in Section 3. The translation performance of the three systems is summarized in Table 4 for the three evaluation criteria WER, PER and BLEU. We observe that the translation quality with the learned segmentation is similar to that with the LDC segmentation. The WER of the system with the learned segmentation is somewhat better, but PER and BLEU are slightly worse. We conclude that it is possible to learn a domain-specific dictionary for Chinese word segmentation from a bilingual corpus. Therefore the translation system is independent of a predefined dictionary, which may be unsuitable for a certain task.

The translation system using no segmentation performs slightly worse. For example, for the WER there is a loss of about 2% relative compared to the system with the LDC segmentation.

Table 4: Translation performance of different segmentation methods (all numbers in percent).

method	error rates		accuracy
	WER	PER	BLEU
no segment.	73.3	56.5	27.6
learned segment.	70.4	54.6	29.1
LDC segment.	71.9	54.4	29.2

5.4 Effect of Segmentation on Translation Results

In this section, we present three examples of the effect that segmentation may have on translation quality. For each of the three examples in

Figure 4, we show the segmented Chinese source sentence using either the LDC dictionary or the self-learned dictionary, the corresponding translation and the human reference translation.

In the first example, the LDC dictionary leads to a correct segmentation, whereas with the learned dictionary the segmentation is erroneous. The second and third token should be combined ("Hong Kong"), whereas the fifth token should be separated ("stabilize in the long term"). In this case, the wrong segmentation of the Chinese source sentence does not result in a wrong translation. A possible reason is that the translation system is based on word groups and can recover from these segmentation errors.

In the second example, the segmentation with the LDC dictionary produces at least one error. The second and third token should be combined ("this"). It is possible to combine the seventh and eighth token to a single word because the eighth token shows only the tense. The segmentation with the learned dictionary is correct. Here, the two segmentations result in different translations.

In the third example, both segmentations are incorrect and these segmentation errors affect the translation results. In the segmentation with the LDC dictionary, the first Chinese characters should be segmented as a separate word. The second and third character and maybe even the fourth character should be combined to one word.[2] The fifth and sixth character should be combined to a single word. In the segmentation with the learned dictionary, the fifth and sixth token (seventh and eighth character) should be combined ("isolated"). We see that this term is missing in the translation. Here, the segmentation errors result in translation errors.

6 Discussion and Future Work

We have presented a new method for Chinese word segmentation. It avoids the use of a predefined dictionary and instead learns a corpus-specific dictionary from the bilingual training corpus.

The idea is extracting a self-learned dictionary from the trained alignment models. This method has the advantage that the word entries in the dictionary all occur in the training data, and its content is much closer to the training text as a predefined dictionary, which can never cover all possible word occurrences. Here, if the content of the test corpus is closer to that of the

[2]This is an example of an ambiguous segmentation.

Example 1

LDC dictionary:
有利 香港 经济 长期 繁荣 稳定 。
It will benefit Hong Kong's economy to prosper and stabilize in the long term.

Learned dictionary:
有利 香 港 经济 长期繁荣 稳定 。
It will benefit Hong Kong's economy to prosper and stabilize in the long term.

Reference:
It will be benificial for the stability and prosperity of Hong Kong in the long run.

Example 2

LDC dictionary:
但是 这 次 会议 呢 还是 取得 了 一定 的 进展 。
but this meeting down or achieved certain progress.

Learned dictionary:
但是 这次 会议 呢 还是 取得了 一定 的 进展 。
however, this meeting straight down still achieved certain progress.

Reference:
Neverless, this meeting has achieved some progress.

Example 3

LDC dictionary:
... 在世 界 上面 临 孤立 的 又 一个 证明 ...
... the unification of the world carried adjacent isolate of proof, ...

Learned dictionary:
... 在 世界 上 面临 孤 立 的 又一个 证明 ...
... in the world faced with a became another proof, ...

Reference:
... another proof that ... is facing isolation in the world ...

Figure 4: Translation examples using the learned dictionary and the LDC dictionary.

training corpus, the quality of the dictionary is higher and the translation performance would be better.

The experiments showed that the translation quality with the learned segmentation is competitive with the LDC segmentation. Additionally, we have shown the feasibility of a Chinese–English statistical machine translation system that works without any word segmentation. There is only a minor loss in translation performance. Further improvements could be possible by tuning the system toward this specific task.

We expect that our method could be improved by considering the word length as discussed in Section 4.3. As shown in the word length statistics, long words with more than four characters occur only occasionally. Most of them are named entity words, which are written in English in upper case. Therefore, we can apply a simple rule: we accept a long Chinese word only if the corresponding English word is in upper case. This should result in an improved dictionary. An alternative way is to use the word length statistics in Table 3 as a prior distribution. In this case, long words would get a penalty, because their prior probability is low.

Because the extraction of our dictionary is based on bilingual information, it might be interesting to combine it with methods that use monolingual information only.

For Chinese–English, there is a large number of bilingual corpora available at the LDC. Therefore using additional corpora, we can expect to get an improved dictionary.

References

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

LDC. 2003a. LDC Chinese resources home page. http://www.ldc.upenn.edu/Projects/Chinese/LDC_ch.htm.

LDC. 2003b. LDC resources home page. http://www.ldc.upenn.edu/Projects/TIDES/mt2004cn.htm.

NIST. 2003. Machine translation home page. http://www.nist.gov/speech/tests/mt/index.htm.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

D. D. Palmer. 1997. A trainable rule-based algorithm for word segmentation. In *Proc. of the 35th Annual Meeting of ACL and 8th Conference of the European Chapter of ACL*, pages 321–328, Madrid, Spain, August.

K. A. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, September.

R. W. Sproat and C. Shih. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Languages*, 4:336–351.

M. Sun, D. Shen, and B. K. Tsou. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proc. of the 36th Annual Meeting of ACL and 17th Int. Conf. on Computational Linguistics (COLING-ACL 98)*, pages 1265–1271, Montreal, Quebec, Canada, August.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.