Improving Word Alignment by Adjusting Chinese Word Segmentation

Ming-Hong Bai[1,2] **Keh-Jiann Chen**[1] **Jason S. Chang**[2]
1 Institute of Information Science, Academia Sinica
2 Department of Computer Science, National Tsing-Hua University
mhbai@sinica.edu.tw kchen@iis.sinica.edu.tw jschang@cs.nthu.edu.tw

Abstract

Most of the current Chinese word alignment tasks often adopt word segmentation systems firstly to identify words. However, word-mismatching problems exist between languages and will degrade the performance of word alignment. In this paper, we propose two unsupervised methods to adjust word segmentation to make the tokens 1-to-1 mapping as many as possible between the corresponding sentences. The first method is learning affix rules from a bilingual terminology bank. The second method is using the concept of impurity measure motivated by the decision tree. Our experiments showed that both of the adjusting methods improve the performance of word alignment significantly.

1 Introduction

Word alignment is an important preprocessing task for statistical machine translation. There have been many statistical word alignment methods proposed since the IBM models have been introduced. Most existing methods treat word tokens as basic alignment units (Brown et al., 1993; Vogel et al., 1996; Deng and Byrne, 2005), however, many languages have no explicit word boundary markers, such as Chinese and Japanese. In these languages, word segmentation (Chen and Liu, 1992; Chen and Bai, 1998; Chen and Ma, 2002; Ma and Chen, 2003; Gao et al., 2005) is often carried out firstly to identify words before word alignment (Wu and Xia, 1994). However, the differences in lexicalization may degrade word alignment performance, for different languages may realize the same concept using different numbers of words (Ma et al., 2007; Wu, 1997). For instance, Chinese multi-syllabic words composed of more than one meaningful morpheme which may be translated to several English words. For example, the Chinese word *教育署* is composed of two meaning units, *教育* and *署*, and is translated to *Department of Education* in English. The morphemes *教育* and *署* have their own meanings and are translated to *Education* and *Department* respectively. The phenomenon of lexicalization mismatch will degrade the performance of word alignment for several reasons. The first reason is that it will reduce the cooccurrence counts of Chinese and English tokens. Consider the previous example. Since *教育署* is treated as a single unit, it does not contribute to the occurrence counts of *Education*/*教育* and *Department*/*署* token pairs. Secondly, the rarely occurring compound word may cause the *garbage collectors* effect (Moore, 2004; Liang et al., 2006), aligning a rare word in source language to too many words in the target language, due to the frequency imbalance with the corresponding translation words in English (Lee, 2004). Finally, the IBM models (Moore, 2004) impose the limitation that each word in the target sentence can be generated by at most one word in the source sentence. In this case, a many-to-one alignment, links a phrase in the source sentence to a single token in the target sentence, is not allowed, forcing most links of a phrase in the source sentence to be abolished. As in the previous example, when aligning from English to Chinese, *教育署* can only be linked to one of the English words, say *Education*, because of the limitation of the IBM model. However for remedy, many of the current word alignment methods combine the results of both alignment directions, via *intersection* or

grow-diag-final heuristic, to improve the alignment reliability (Koehn et al., 2003; Liang et al., 2006; Ayan et al., 2006; DeNero et al., 2007). However the many-to-one link limitation will undermine the reliability due to the fact that some links are not allowed in one of the directions.

In this paper, we propose two novel methods to adjust word segmentation so as to decrease the effect of lexicalization differences to improve word alignment performance. The main idea of our methods is to adjust Chinese word segmentation according to their translation derived from parallel sentences in order to make the tokens compatible to 1-to-1 mapping between the corresponding sentences. The first method is based on learning a set of affix rules from bilingual terminology bank, and adjusting the segmentation according to these affix rules when preprocessing the Chinese part of the parallel corpus. The second method is based on the so-called *impurity* measure, which was motivated by the decision tree (Duda et al., 2001).

2 Related Works

Our methods are motivated by the translation-driven segmentation method proposed by Wu (1997) to segment words in a way to improve word alignment. However, Wu's method needs a translation lexicon to filter out the links which were not in the lexicon and the result was only evaluated on the sentence pairs which were covered by the lexicon.

A word packing method has been proposed by Ma et al. (2007) to improve the word alignment task. Before carrying out word alignment, this method packs several consecutive words together when those words believed to correspond to a single word in the other language. Our basic idea is similar to this, but on the contrary, we try to unpack words which are translations of several words in the other language. Since the word packing method treats the packed consecutive words as a single token, as we mentioned in the previous section, it weakens the association strength of translation pairs of their morphemes while applying the IBM word alignment model.

A lot of morphological analysis methods have been proposed to improve the performance of word alignment for inflectional language (Lee et al., 2003; Lee, 2004; Goldwater, 2005). They proposed to split a word into a morpheme sequence of the pattern prefix*-stem-suffix* (* denotes zero or more occurrences of a morpheme). Their experiments showed that morphological analysis can improve the quality of machine translation by reducing data sparseness and by making the tokens in two languages correspond more 1-to-1. However, these segmentation methods were developed from the monolingual perspective.

3 Adjusting Word Segmentation

The goal of word segmentation adjustment is to adjust the segmentation of Chinese words such that we have as many 1-to-1 links to the English words as possible. In this task, we will face the problem of finding the proper morpheme boundaries for Chinese words. The challenge is that almost all characters of Chinese are morphemes and therefore almost every character boundary in a word could be the boundary of a morpheme, there is no simple rules to find the suitable boundaries of morphemes. Furthermore, not all meaningful morphemes need to be segmented to meet the requirement of 1-to-1 mapping. For example, *washing machine/洗衣機* can be segmented into *洗衣* and *機* corresponding to *washing* and *machine* while *heater/暖氣機* does not need, it depends on their translations.

In this paper, we have proposed two different methods to solve this problem: 1. learning affix rules from terminology bank to segment morphemes and 2. using *impurity* measure to finding the morpheme boundaries. The detail of these methods will be described in the following sections.

4 Affix Rule Method

The main idea of this method is to segment a Chinese word according to some properly designed conditional dependent affix rules. As shown in Figure 1, each rule is composed of three conditional constraints, a) affix condition, b) English word condition and c) exception condition. In the affix condition, we place a underscore on the left of a morpheme, such as *_機*, to denote a suffix and on the right, such as *副_*, to denote a prefix. The affix rules are applied to each word by checking the following three conditions:

1. The target word has the affix.

2. The English word which is the target of translation exists in the parallel sentence.
3. The target word does not contain the morphemes in the exception list (The morpheme in the exception list shows an alternative segmentation.).

If the target word satisfies all of the above conditions of any rule, then the morpheme should be separated from the word. The remaining problem will be how to derive the set of affix rules.

affix	English word	exception
_機	machine	
_機	engine	
副_	vice	
副_	deputy	副手
_業	industry	工業

Figure 1. Samples of affix rules.

4.1 Training Data

We use an unsupervised method to extract affix rules from a Chinese-English terminology bank[1]. The bilingual terminology bank a total of 1,046,058 English terms with Chinese translations in 63 categories. Among them, 60% or 629,352 terms are compounds. We take the advantage of the terminology bank, that all terminologies are 1-to-1 well translated, to find the best morpheme segmentation from ambiguous segmentations of a Chinese word according to its English counterpart. Then we extracted affix rules from the word-to-morpheme alignment results of terms and translation.

4.2 Word-to-Morpheme Alignment

The training phase of word-to-morpheme alignment is based loosely on word-to-word alignment of the IBM model 1. Instead of using Chinese words, we considered all the possible morphemes. For example, consider the task of aligning *Department of Education* and *教育署* as shown as Figure 2. We use the EM algorithm to train the translation probabilities of word-morpheme pairs based on IBM model 1.

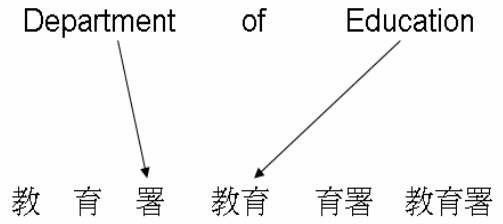


Figure 2. Example of word-to-morpheme alignment.

In the aligning phase, the original IBM model 1 does not work properly as we expected. Because the English words prefer to link to single character and it results that some correct Chinese translations will not be linked. The reason is that the probability of a morpheme, say $p(教育|education)$, is always less than its substring, $p(教|education)$, since whatever *教育* occurs *教* and *育* always occur but not vice versa. So the aligning result will be *教/Education* and *署/Department*, *育* is abandoned. To overcome this problem, a constraint of alignment is imposed to the model to ensure that the aligning result covers every Chinese characters of a target word and no overlapped characters in the result morpheme sequence. For instances, both *教/Education 署/Department* and *教育/Education 育署/Department* are not allowed alignment sequences. The constraint is applied to each possible aligning result. If the alignment violates the constraint, it will be rejected.

Since the new alignment algorithm must enumerate all of the possible alignments, the process is very time consuming. Therefore, it is advantageous to use a bilingual terminology bank rather than a parallel corpus. The average length of terminologies is short and much shorter than a typical sentence in a parallel corpus. This makes words to morphemes alignment computationally feasible and the results highly accurate (Chang et al., 2001; Bai et al., 2006). This makes it possible to use the result as pseudo gold standards to evaluate affix rules as described in section 4.3.

1 The bilingual terminology bank was compiled by the National Institute for Compilation and Translation. It is freely download at http://terms.nict.gov.tw by registering your information.

air|空氣 refrigeration|冷凍 machine|機
building|建築 industry|業
compound|複式 steam|蒸汽 engine|機
electronics|電子 industry|業
vice|副 chancellor|校長

Figure 3. Sample of word-to-morpheme alignment.

4.3 Rule Extraction

After the alignment task, we will get a word-to-morpheme aligned terminology bank as shown in Figure 3. We can subsequently extract affix rules from the aligned terminology bank by the following steps:

1) **Generate candidates of affix rule:**
   For each alignment, we produce all alignment links as affix rules. For instance, with (*electronics*| *電 子* *industry*| *業*), we would produce two rules:

   (a) *電子_, electronics*
   (b) *_業, industry*

2) **Evaluate the rules:**
   The precision of each candidate rule is estimated by applying the rule to segment the Chinese terms. If a Chinese term contains the affix shown in the rule, the affix will be segmented. The results of segmentation are then to compare with the segmentation results of the alignments done by the algorithm of the section 4.2 as pseudo gold standards. Some example results of rule evaluations are shown in Figure 4.

affix	English word	Rule Applied	Correct segments	precision
主_	master	458	378	0.825
週期_	periodic	130	100	0.769
視訊_	video	46	40	0.870
_鍊	chain	147	107	0.728
_箱	box	716	545	0.761

Figure 4. Sample evaluations of candidate rules.

3) **Adding exception condition:**
   In the third step, we sort the rules according to their precision rates in descending order, resulting in rules $R_1..R_n$. And then for each R_i, we scan R_1 to R_{i-1}, if there is a rule, R_j, have the same English word condition and the affix condition of R_i subsume that of R_j, then we add affix condition of R_j as exception condition of R_i. For example, *_業, industry* and *_工業, industry* are rule candidates in the sorted table and have the same English word condition. Furthermore, the condition *_業* subsumes that of *工業*, we add *工業* to the exception condition of the rule with a shorter affix.

4) **Reevaluate the rules with exception condition:**
   After adding the exception conditions, the rules are reevaluated with considering the exception condition to get new evaluation scores.

5) **Select rules by scores**:
   Finally, filter out the rules with scores lower than a threshold[2].

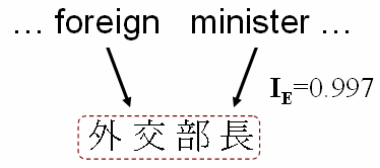
The reason of using exception condition is that an affix is usually an abbreviation of a word, such as *_業* is an abbreviation of *工業*. In general, a full morpheme is preferred to be segmented than its abbreviation while both occurred in a target word. For example, when applying rules to *電子工業 /electronic industry*, *_工業,industry* is preferred than *_業,industry*. However, in the evaluation step, precision rate of *_業,industry* will be reduced when applying to full morphemes, such as *電子工業 /electronic industry*, and then could be filtered out if the precision is lower than the threshold.

5 Impurity Measure Method

The impurity measure was used by decision tree (Duda et al., 2001) to split the training examples into smaller and smaller subsets progressively according to features and hope that all the samples in each subset is as *pure* as possible. For convenient, they define the *impurity* function rather than the *purity* function of a subset as follows:

$$impurity(S) = -\sum_j P(w_j)\log_2 P(w_j)$$

[2] We set the threshold as 0.7.



(a) impurity value of *外交部長*.

(b) impurity values of *外交* and *部長*.

Figure 5. Examples of impurity values.

Where $P(w_j)$ is the fraction of examples at set S that are in category w_j. By the well-known properties of entropy if all the examples are of the same category the impurity is 0; otherwise it is positive, with the greatest value occurring when the different classes are equal likely.

5.1 Impurity Measure of Translation

In our experiment, the impurity measure is used to split a Chinese word into two substrings and hope that all the characters in a substring are generated by the parallel English words as *pure* as possible. Here, we treat a Chinese word as a set of characters, the parallel English words as categories and the fraction of examples is redefined by the expected fraction number of characters that are generated by each English word. So we redefine the *entropy impurity* as follows:

$$I_E(f;\mathbf{e},\mathbf{f}) = -\sum_{\forall e \in \mathbf{e}} c(f \mid e;\mathbf{e},\mathbf{f}) \log_2 c(f \mid e;\mathbf{e},\mathbf{f})$$

In which f denotes the target Chinese word, **e** and **f** denote the parallel English and Chinese sentence that f belongs to and $c(f \mid e;\mathbf{e},\mathbf{f})$ is the expected fraction number of characters in f that are generated by word e. The expected fraction number can be defined as follows:

$$c(f \mid e;\mathbf{e},\mathbf{f}) = \frac{\sum_{\forall c \in f} p(c \mid e)}{\sum_{\forall e \in \mathbf{e}} \sum_{\forall c \in f} p(c \mid e)}$$

Where $p(c \mid e)$ denotes the translation probability of Chinese character c given English word e.

For example, as shown in Figure 5, the impurity value of *外交部長*, Figure 5.(a), is much higher than values of *外交* and *部長*, Figure 5.(b). Which means that the generating relations from English to Chinese tokens are purified by breaking *外交部長* into *外交* and *部長*.

The translation probabilities between Chinese characters and English word can be trained using IBM model 1 by treating Chinese characters as tokens.

5.2 Target Word Selection

In this experiment, we treat the Chinese words which can be segmented into morphemes and linked to different English words as target words. In order to speedup our impurity method only target words will be segmented during the process. Therefore we investigate the actual distribution of target words first, we have tagged 1,573 Chinese words manually with *target* and *non-target*. It turns out that only 6.87% of the Chinese words are tagged as *target* and 94.4% of target words are nouns. The results show that most of the Chinese words do not need to be re-segmented and their POS distribution is very unbalanced. The results show that we can filter out the *non-target* words by simple clues. In our experiment, we use three features to filter out *non-target* words:

1) POS: Since 94.4% of the target words are nouns, we focus our experiment on nouns and filter out words with other POS.
2) One-to-many alignment in GIZA++: Only Chinese words which are linked to multiple English words in the result of GIZA++ are considered to be target words.
3) Impurity measure: the target words are expected to have high impurity values. So the words with a impurity values larger than a threshold are selected as target words[3].

[3] In our experiment, we use 0.3 as our threshold.

5.3 Best Breaking Point

The goal of segmentation adjustment using impurity is to find the best breaking point of a Chinese word according to parallel English words. When a word is broken into two substrings, the new substrings can be compared to original word by the *information gain* which is defined in terms of impurity as follows:

$$IG(f, f_1^i, f_{i+1}^n) = I_E(f;\mathbf{e},\mathbf{f}) - \frac{1}{2} I_E(f_1^i;\mathbf{e},\mathbf{f}) - \frac{1}{2} I_E(f_{i+1}^n;\mathbf{e},\mathbf{f})$$

Where i denotes a break point in f, f_1^i denotes first i characters of f, and f_{i+1}^n denotes last n-i characters of f. If the information gain of a breaking point is positive, the result substrings are considered to be better, i.e. more pure than original word.

The goal of finding the best breaking point can be achieved by finding the point which maximizes the information gain as the following formula:

$$\arg\max_{1\le i<n} IG(f, f_1^i, f_{i+1}^n)$$

Note that a word can be separated into two substrings each time. If we want to segment a complex word composed of many morphemes, just split the word again and again like the construction of decision tree, until the information gain is negative or less than a threshold[4].

6 Experiments

In order to evaluate the effect of our methods on the word alignment task, we preprocessed parallel corpus in three ways: First we use a state-of-the-art word segmenter to tokenize the Chinese part of the corpus. Then, we used the affix rules to adjust word segmentation. Finally, we do the same but by using the impurity measure method. We used the GIZA++ package (Och and Ney, 2003) as the word alignment tool to align tokens on the three copies of preprocessed parallel corpora.

We used the first 100,000 sentences of Hong Kong News parallel corpus from LDC as our training data. And 112 randomly selected parallel sentences were aligned manually with *sure* and *possible* tags, as described in (Och and Ney, 2000), and we used these annotated data as our gold standard in testing.

Because of the modification of Chinese tokens caused by the word segmentation adjustment, a problem has been created when we wanted to compare the results to the copy which did not undergo adjustment. Therefore, after the alignment was done, we merged the alignment links related to tokens that were split up during adjustment. For example, the two links of *foreign*/外交 *minister*/部長 were merged as *foreign minister*/外交部長.

The evaluation of word alignment results are shown in Table 1, including *precision-recall* and *AER* evaluation methods. In which the *baseline* is alignment result of the unadjusted data. The table shows that after the adjustment of word segmentation, both methods obtain significant improvement over the *baseline*, especially for the English-Chinese direction and the intersection results of both directions. The *impurity* method in particular improves alignment in both English-Chinese and Chinese-English directions.

The improvement of intersection of both directions is important for machine translation. Because the intersection result has higher precision, a lot of machine translation method relies on intersecting the alignment results. The phrase-based machine translation (Koehn et al., 2003) uses the *grow-diag-final* heuristic to extend the word alignment to phrase alignment by using the intersection result. Liang (Liang et al., 2006) has proposed a symmetric word alignment model that merges two simple asymmetric models into a symmetric model by maximizing a combination of likelihood and agreement between the models. This method uses the intersection as the agreement of both models in the training time. The method has reduced the alignment error significantly over the traditional asymmetric models.

In order to analyze the adjustment results, we also manually segment and link the words of Chinese sentences to make the alignments 1-to-1 mapping as many as possible according to their translations for the 112 gold standard sentences. Table 2 shows the results of our analysis, the performance of impurity measure method is also slightly better than the affix rules in both recall and precision measure.

[4] In our experiment, we set 0 as the threshold.

	direction	Recall	precision	F-score	AER
baseline	English-Chinese	68.3	61.2	64.6	35.7
	Chinese-English	79.6	67.0	72.8	27.8
	intersection	59.9	92.0	72.6	26.6
affix rules	English-Chinese	78.2	64.6	70.8	29.8
	Chinese-English	80.2	68.0	73.6	27.0
	intersection	69.1	92.3	79.0	20.2
impurity	English-Chinese	78.1	64.9	70.9	29.7
	Chinese-English	81.4	70.4	75.5	25.0
	intersection	70.2	91.9	79.6	19.8

Table 1. Alignment results based on the standard word segmentation data.

	recall	precision
affix rules	82.35	66.66
impurity	84.31	67.72

Table 2. Alignment results based on the manual word segmentation data.

7 Conclusion

In this paper, we have proposed two Chinese word segmentation adjustment methods to improve word alignment. The first method uses the affix rules learned from a bilingual terminology bank and then applies the rules to the parallel corpus to split the compound Chinese words into morphemes according to its counterpart parallel sentence. The second method uses the impurity method, which was motivated by the method of decision tree. The experimental results show that both methods lead to significant improvement in word alignment performance.

Acknowledgements: This research was supported in part by the National Science Council of Taiwan under NSC Grants: NSC95-2422-H-001-031.

References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT. In *Proceedings of ACL 2006*, pages 9-16, Sydney, Australia.

Ming-Hong Bai, Keh-Jiann Chen and Jason S. Chang. 2006. Sense Extraction and Disambiguation for Chinese Words from Bilingual Terminology Bank. *Computational Linguistics and Chinese Language Processing*, 11(3):223-244.

Petter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311.

Jason S Chang, David Yu, Chun-Jun Lee. 2001. Statistical Translation Model for Phrases(in Chinese). *Computational Linguistics and Chinese Language Processing*, 6(2):43-64.

Keh-Jiann Chen, Ming-Hong Bai. 1998. Unknown Word Detection for Chinese by a Corpus-based Learning Method. *International Journal of Computational linguistics and Chinese Language Processing*, 1998, Vol.3, #1, pages 27-44.

Keh-Jiann Chen, Wei-Yun Ma. 2002. Unknown Word Extraction for Chinese Documents. In *Proceedings of COLING 2002*, pages 169-175, Taipei, Taiwan.

Keh-Jiann Chen, Shing-Huan Liu. 1992. Word Identification for Mandarin Chinese Sentences. In *Proceedings of 14th COLING*, pages 101-107.

John DeNero, Dan Klein. 2007. Tailoring Word Alignments to Syntactic Machine Translation. In *Proceedings of ACL 2007*, pages 17-24, Prague, Czech Republic.

Yonggang Deng, William Byrne. 2005. HMM word and phrase alignment for statistical machine translation. In *Proceedings of HLT-EMNLP 2005*, pages 169-176, Vancouver, Canada.

Richard O. Duda, Peter E. Hart, David G. Stork. 2001. *Pattern Classification*. John Wiley & Sons, Inc.

Jianfeng Gao, Mu Li, Andi Wu and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: a pragmatic approach. *Computational Linguistics,* 31(4)

Sharon Goldwater, David McClosky. 2005. Improving Statistical MT through Morphological Analysis. In

Proceedings of HLT/EMNLP 2005, pages 676-683, Vancouver, Canada.

Philipp Koehn, Franz J. Och, Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT/NAACL 2003*, pages 48-54, Edmonton, Canada.

Young-Suk Lee. 2004. Morphological Analysis for Statistical Machine Translation. In *Proceedings of HLT-NAACL 2004*, pages 57-60, Boston, USA.

Young-Suk Lee, Kishore Papineni, Salim Roukos. 2003. Language Model Based Arabic Word Segmentation. In *Proceedings of ACL 2003*, pages 399-406, Sapporo, Japan.

Percy Liang, Ben Taskar, Dan Klein. 2006. Alignment by Agreement. In *Proceedings of HLT-NAACL 2006*, pages 104-111, New York, USA.

Wei-Yun Ma, Keh-Jiann Chen. 2003. A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. In *Proceedings of ACL 2003, Second SIGHAN Workshop on Chinese Language Processing*, pp31-38, Sapporo, Japan.

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

Robert C. Moore. 2004. Improving IBM Word-Alignment Model 1. In *Proceedings of ACL 2004*, pages 519-526, Barcelona, Spain.

Franz Josef Och, Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

Franz J. Och, Hermann Ney., Improved Statistical Alignment Models, In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, Hong Kong, pp. 440-447.

Stefan Vogel, Hermann Ney, Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING 1996*, pages 836-841, Copenhagen, Denmark.

Dekai Wu, Xuanyin Xia. 1994. Learning an English-Chinese Lexicon from a Parallel Corpus. In *Proceedings of AMTA 1994*, pages 206-213, Columbia, MD.

Dekai Wu. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics*, 23(3):377-403.