Improving Word Alignment by Adjusting Chinese Word Segmentation

Ming-Hong Bai
mhbai@sinica.edu.tw

Keh-Jiann Chen
kchen@iis.sinica.edu.tw

Jason S. Chang
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 supporting task for statistical machine translation. There are many statistical word alignment methods have been proposed since the IBM models have been introduced. Most of the current 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 etc. In these languages, word segmentation systems (Chen and Liu, 1992; Chen and Bai, 1998; Chen and Ma, 2002; Ma and Chen, 2003; Gao et al., 2005) are often adopted firstly to identify words before word alignment (Wu and Xia, 1994). However, it will cause a mismatching problem since different languages may realize the same concept using varying numbers of words (Ma et al., 2007; Wu, 1997). In Chinese language, multi-syllabic words are composed of more than one meaningful morpheme which may be translated to English words individually. For example, the Chinese word *教育署* is composed of two morphemes, *教育* and *署*, and its English translation, *Department of Education*, is composed of three words, the morphemes *教育* and *署* have their own meanings and are translated to *Education* and *Department* respectively. The phenomenon of word mismatching will degrade the performance of word alignment for some reasons. The first reason is that it will reduce the collocation frequency of Chinese and English tokens. Consider the previous example. Since *教育署* is treated as a single unit, it has no collocation contribution to the *Education*/*教育* and *Department*/*署* token pairs in this case. The second reason is the rarely occurring compound word may cause the *garbage collectors* problem (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). The third reason is the limitation of the IBM models (Moore, 2004), 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, most links of a phrase in the source sentence are forced to be abolished.

In this paper, we have proposed two novel methods to adjust word segmentation for word alignment to avoid the word-mismatching problem. The main idea of our methods are adjusting Chinese word segmentation according to their translation parallel sentences in order to make the tokens 1-to-1 mapping between the corresponding sentences. The first method is using the prefix and suffix rules, learning from bilingual terminology

bank, to adjust the segmentation of the testing data. The second method is using the *impurity* measure, which was motivated by the decision tree (Duda et al., 2001), to adjust the segmentation of the testing data.

2 Related Works

Our methods are motivated by the translation-driven segmentation method proposed by Wu (Wu, 1997) to address the word segmentation problem of word alignment for Chinese. 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 pairs which were listed in the lexicon.

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 separate 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 because of reducing data sparseness and increasing similarity between languages. However, their segmenter was trained in monolingual without considering bilingual correspondence. Hence didn't really solve the problem of word mismatching.

3 Word Segmentation Adjustment

The goal of word segmentation adjustment is to adjust the segmentation of Chinese words into morphemes which makes the 1-to-1 links to the English words as many 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 prefix and suffix 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) prefix-suffix condition, b) English word condition and c) exception condition. In the prefix-suffix 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 prefix or suffix.
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.

prefix-suffix	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 Rule Extraction

We use an unsupervised method to extract affix rules from a Chinese-English terminology bank. The bilingual terminology bank contains 63 classes of terminologies, a total of 1,046,058 Chinese terms with their English translations. Among them, 629,352 terms are compounds,

which is about 60 percent of all. We took 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 translation. Then extracting affix rules from the word-to-morpheme aligned terminologies.

The training phase of word-to-morpheme alignment is by a modified word-to-word alignment of the IBM model 1. The modification is that we list all the possible morpheme candidates instead of words on the target language, and place English words on the source part as usual. Here is an example of aligning *Department of Education* to *教育署* as shown as figure 2. We use the EM algorithm of IBM model 1 to train the translation probabilities of word-morpheme pairs.

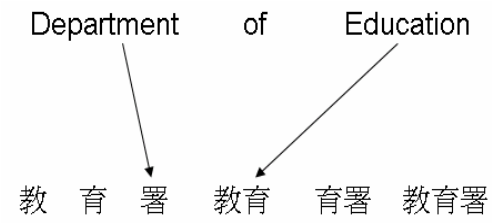


Figure 2. Example of word-to-morpheme alignment.

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 wise 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 should cover every Chinese characters of a target word and no overlapped characters in the result morpheme sequence, such as *教育* and *育署*, are not allowed morpheme sequence. 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, that is the reason why we must use parallel terminology bank rather than parallel corpus. The average length of terminologies is short and much shorter than a sentence. This makes words to morphemes alignment computationally feasible and result accurate.

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. Samples of word-to-morpheme alignment.

After the alignment task, we will get a word-to-morpheme aligned corpus as exemplified in Figure 3. Then we can extract rules from aligned corpus by the following steps:

1. Extracting rule candidates:
   List prefix or suffix with its translation for each alignment as the following example:
   electronics|*電子* *industry*|*業*→
   *電子_*, *electronics*
   _業, *industry*
2. Evaluate the rules:
   Apply each candidate rule to the original terminology bank to separate prefix and suffix, and then use the alignments as correct answers to evaluate the segmentation results.
3. Adding exception condition:
   Sort the rules according to their accuracy 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 prefix-suffix condition of R_i is covered (substring) by R_j, then we add prefix-suffix condition of R_j as exception condition of R_i.
4. Reevaluate the rules with exception condition:
   After adding the exception conditions, the rules are evaluated again to get the new evaluation scores.

The reason of using exception condition is that the prefix or suffix is usually an abbreviation of a word. In general, a full morpheme is prefer to be segmented than it abbreviation while both occurred in a target word. For example, the suffix *_業* is the abbreviation of *工業*, the correct rates of rule _業

/industry will be reduced when applying to the words with suffix *_工業*.

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:

$$impurity(S) = -\sum_{j} P(w_j)\log_2 P(w_j)$$

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.

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:

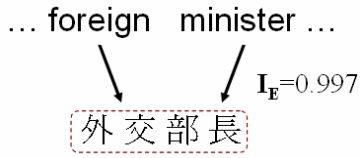
$$I_E(f;\mathbf{e},\mathbf{f}) = -\sum_{\forall e\in\mathbf{e}} c(f \mid e;\mathbf{e},\mathbf{f})\log_2 c(f \mid e;\mathbf{e},\mathbf{f})$$

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:

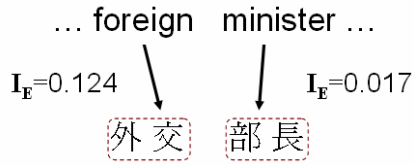
$$c(f \mid e;\mathbf{e},\mathbf{f}) = \frac{\sum_{\forall c\in f} p(c \mid e)}{\sum_{\forall e\in\mathbf{e}}\sum_{\forall c\in f} p(c \mid e)}$$

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

For example, as shown in figure 4, the impurity value of 外交部長, figure (a), is much higher than values of 外交 and 部長, figure (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*.



(a) impurity value of *外交部長*.



(b) impurity values of *外交* and *部長*.

Figure 4. Examples of impurity values.

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

5.1 Segmentation

The goal of segmentation 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 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.

6 Experiments

In order to evaluate the influence of our methods on the word alignment task, we prepare tree copies of testing data, the first copy is the original data which has been preprocessed by a standard word segmentation system, the second copy is adjusted the word segmentation by the affix rules, the third copy is adjusted the word segmentation by the impurity method. We use the GIZA++ package (Och and Ney, 2003) as the word alignment tool to align tokens on the tree copies.

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

Because the word segmentation adjustment will modify the Chinese tokens, it is not possible to evaluate the word alignment results directly. After the alignment task, we have to merge the tokens which belong to the same word in the original data. The merging task will merge both the tokens and their alignment links. For example, the Chinese tokens 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 original data. The table shows that after the adjustment of word segmentation, both methods obtain significant improvement over the *baseline*, especially the English-Chinese direction and the intersection results of both directions. The *impurity* method also improves the Chinese-English direction.

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 intersection 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 maximize a combination of data 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 analysis 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 method is also slightly better than the affix rules in both recall and precision measure.

	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 best segmentation adjusted data.

7 Conclusion

In this paper, we have proposed two Chinese word segment adjustment methods to improve the word alignment. The first method uses the affix rules extracted from a bilingual terminology bank. And then apply the rules to the testing data 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, to achieve the same goal. The experimental results show that both of the methods have significant improvement to the word alignment.

References

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.

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.

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. 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.