Improving Domain-Specific Word Alignment with a General Bilingual Corpus

WU Hua, WANG Haifeng
Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza, No.1, East Chang An Ave., Dong Cheng District
Beijing, 100738, China
{wuhua, wanghaifeng}@rdc.toshiba.com.cn

Abstract. In conventional word alignment methods, some employ statistical models or statistical measures, which need large-scale bilingual sentence-aligned training corpora. Others employ dictionaries to guide alignment selection. However, these methods achieve unsatisfactory alignment results when performing word alignment on a small-scale domain-specific bilingual corpus without terminological lexicons. This paper proposes an approach to improve word alignment in a specific domain, in which only a small-scale domain-specific corpus is available, by adapting the word alignment information in the general domain to the specific domain. This approach first trains two statistical word alignment models with the large-scale corpus in the general domain and the small-scale corpus in the specific domain respectively, and then improves the domain-specific word alignment with these two models. Experimental results show a significant improvement in terms of both alignment precision and recall, achieving a relative error rate reduction of 21.96% as compared with state-of-the-art technologies.

1 Introduction

Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) [3]. Besides being used in SMT, it is also used in translation lexicon building [8], transfer rule learning [9], example-based machine translation [13], translation memory systems [12], etc.

In previous alignment methods, some researchers modeled the alignments as hidden parameters in a statistical translation model [3], [10] or directly modeled them given the sentence pairs [4]. Some researchers use similarity and association measures to build alignment links [1], [11], [14]. In addition, Wu [15] used a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments. However, All of these methods require a large-scale bilingual corpus for training. When the large-scale bilingual corpus is not available, some researchers use existing dictionaries to improve word alignment [6]. However, few works address the problem of domain-specific word alignment when neither the large-scale domain-specific bilingual corpus nor the domain-specific translation dictionary is available.

In this paper, we address the problem of word alignment in a specific domain, in which only a small-scale corpus is available. In the domain-specific corpus, there are two kinds of words. Some are general words, which are also frequently used in the general domain. Others are domain-specific words, which only occur in the specific domain. In general, it is not quite hard to obtain a large-scale general bilingual corpus while the available domain-specific bilingual corpus is usually quite small. Thus, we use the bilingual corpus in the general domain to improve word alignments for general words and the bilingual corpus in the specific domain for domain-specific words. In other words, we will adapt the word alignment information in the general domain to the specific domain.

Although the adaptation technology is widely used for other tasks such as language modeling, few literatures, to the best of our knowledge, directly address word alignment adaptation. The work most closely related to ours is the statistical translation adaptation described in [7]. Langlais used terminological lexicons to improve the performance of a statistical translation engine, which is trained on a general bilingual corpus and used to translate a manual for military snipers. The experimental results showed that this adaptation method could reduce word error rate on the translation task.

In this paper, we perform word alignment adaptation from the general domain to a specific domain (in this study, a user manual for a medical system) with four steps. (1) We train a word alignment model using a bilingual corpus in the general domain; (2) We train another word alignment model using a small-scale bilingual corpus in the specific domain; (3) We build two translation dictionaries according to the alignment results in (1) and (2) respectively; (4) For each sentence pair in the specific domain, we use the two models to get different word alignment results and improve the results according to the translation dictionaries. Experimental results show that our approach improves domain-specific word alignment in terms of both precision and recall, achieving a 21.96% relative error rate reduction.

The remainder of the paper is organized as follows. Section 2 introduces the statistical word alignment method and analyzes the problems existing in this method for the domain-specific task. Section 3 describes our word alignment adaptation algorithm. Section 4 describes the evaluation results. The last section concludes our approach and presents the future work.

2 Statistical Word Alignment

In this section, we apply the IBM statistical word alignment models to our domain-specific corpus and analyze the alignment results. The tool used for statistical word alignment is GIZA++ [10]. With this tool, we compare the word alignment results of three methods. These methods use different corpora to train IBM word alignment model 4. The method “G+S” directly combines the bilingual sentence pairs in the general domain and in the specific domain as training data. The method “G” only uses the bilingual sentence pairs in the general domain as training data. The method “S” only uses the bilingual sentence pairs in the specific domain as training data.

2.1 Training and Testing Data

We have a sentence aligned English-Chinese bilingual corpus in the general domain, which includes 320,000 bilingual sentence pairs, and a sentence aligned English-Chinese bilingual corpus in the specific domain (a user manual for a medical system), which includes 546 bilingual sentence pairs. From this domain-specific corpus, we randomly select 180 pairs as testing data. The remained 366 pairs are used as domain-specific training data. [1]

The Chinese sentences in both the training set and the testing set are automatically segmented into words. Thus, there are two kinds of errors for word alignment: one is the word segmentation error and the other is the alignment error. In Chinese, if a word is incorrectly segmented, the alignment result is also incorrect. For example, for the Chinese sentence "诊断床面的警告标签" (Warning label for the couch-top), our system segments it into "诊断/床/面的/警告/标签". The sequence "床面的" is incorrectly segmented into "床/面的(couch/taxi)", which should be "床面/的(couch-top/of)". Thus, the segmentation errors in Chinese may change the word meaning, which in turn cause alignment errors.

In order to exclude the effect of the segmentation errors on our alignment results, we correct the segmentation errors in our testing set. The alignments in the testing set are manually annotated, which includes 1,478 alignment links.

2.2 Overall Performance

There are several different evaluation methods for word alignment [2]. In our evaluation, we use evaluation metrics similar to those in [10]. However, we do not classify alignment links into sure links and possible links. We consider each alignment (s,t) as a sure link, where both *s* and *t* can be words or multi-word units.

If we use S_G to represent the alignments identified by the proposed methods and S_C to denote the reference alignments, the methods to calculate the precision, recall, and f-measure are shown in Equation (1), (2) and (3). According to the definition of the alignment error rate (AER) in [10], AER can be calculated with Equation (4). Thus, the higher the f-measure is, the lower the alignment error rate is.

$$precision = \frac{|S_G \cap S_C|}{|S_G|} \quad (1)$$

$$recall = \frac{|S_G \cap S_C|}{|S_C|} \quad (2)$$

$$fmeasure = \frac{2*|S_G \cap S_C|}{|S_G| + |S_C|} \quad (3)$$

$$AER = 1 - \frac{2*|S_G \cap S_C|}{|S_G| + |S_C|} = 1 - fmeasure \quad (4)$$

[1] Generally, a user manual only includes several hundred sentences.

With the above metrics, we evaluate the three methods on the testing set with Chinese as the source language and English as the target language. The results are shown in Table 1. It can be seen that although the method "G+S" achieves the best results among others, it performs just a little better than the method "G". This indicates that adding the small-scale domain-specific training sentence pairs into the general corpus doesn't greatly improve the alignment performance.

Table 1. Statistical Word Alignment Results

Method	Precision	Recall	AER
G+S	0.7140	0.6942	0.2961
G	0.7136	0.6847	0.3014
S	0.4486	0.4066	0.5735

2.3 Result Analysis

We use A, B and C to represent the set of correct alignment links extracted by the method "G+S", the method "G" and the method "S", respectively. From the experiments, we get $|A|=1026$, $|B|=1012$ and $|C|=601$ and get two intersection sets $|D|=|A\cap C|=524$ and $|E|=|B\cap C|=516$. Thus, about 14% alignment links of C are not covered by B. That is to say, although the size of the domain-specific corpus is very small, it can produce word alignment links that are not covered by the general corpus. These alignment links usually include domain-specific words. Moreover, about 13% alignment links of C are not covered by A. This indicates that, by combining the two corpora, the method "G+S" still cannot detect the domain-specific alignment links. At the same time, about 49% of alignment links in both A and B are not covered by the set C.

For example, in the sentence pair in Figure 1, there is a domain-specific word "multislice". For this word, both the method "G+S" and "G" produce a wrong alignment link (multislice, 扫描) while the method "S" produces a correct word alignment link (multislice, 多扫描层). However, the general word alignment link (refer to, 参见) is detected by both the method "G+S" and the method "G" but not detected by the method "S".

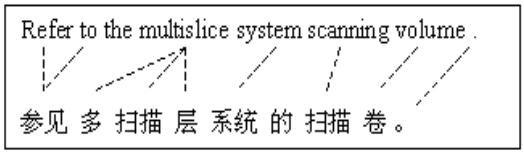

Fig. 1. Alignment Example

Based on the above analysis, it can be seen that it is not effective to directly combine the bilingual corpus in the general domain and in the specific domain as training data. However, the correct alignment links extracted by the method "G" and

those extracted by the method "S" are complementary to each other. Thus, we can develop a method to improve the domain-specific word alignment based on the results of both the method "G" and the method "S".

Another kind of errors is about the multi-word alignment links[2]. The IBM statistical word alignment model only allows one-to-one or more-to-one alignment links. However, the domain-specific terms are usually aligned to more than one Chinese word. Thus, the multi-word unit in the corpus cannot be correctly aligned using this statistical model. For this case, we will use translation dictionaries as guides to modify some alignment links and get multi-word alignments.

3 Word Alignment Adaptation

According to the result analysis in Section 2.3, we take two measures to improve the word alignment results. One is to combine the word alignment results of both the method "G" and the method "S". The other is to use translation dictionaries.

3.1 Bi-directional Word Alignment

In statistical translation models [3], only one-to-one and more-to-one word alignment links can be found. Thus, some multi-word units cannot be correctly aligned. In order to deal with this problem, we perform translation in two directions (English to Chinese, and Chinese to English) as described in [10]. The GIZA++ toolkit is used to perform statistical word alignment.

For the general domain, we use SG_1 and SG_2 to represent the alignment sets obtained with English as the source language and Chinese as the target language or vice versa. For alignment links in both sets, we use *i* for English words and *j* for Chinese words.

$$SG_1 = \{(A_j, j) \mid A_j = \{a_j\}, a_j \geq 0\} \quad \textbf{(5)}$$

$$SG_2 = \{(i, A_i) \mid A_i = \{a_i\},\ a_i \geq 0\} \quad \textbf{(6)}$$

Where, $a_x (x = i, j)$ represents the index position of the source word aligned to the target word in position *x*. For example, if a Chinese word in position *j* is connected to an English word in position *i*, then $a_j = i$. If a Chinese word in position *j* is connected to English words in positions i_1 and i_2, then $A_j = \{i_1, i_2\}$.

Based on the two alignment sets, we obtain their intersection set, union set[3] and subtraction set.

---

[2] Multi-word alignment links means one or more source words aligned to more than one target word or vice versa.

[3] In this paper, the union operation does not remove the replicated elements. For example, if set one includes two elements {1, 2} and set two includes two elements {1, 3}, then the union of these two sets becomes {1, 1, 2, 3}.

Intersection: $SG = SG_1 \cap SG_2$

Union: $PG = SG_1 \cup SG_2$

Subtraction: $MG = \mathrm{PG} - 2 * SG$

Thus, the subtraction set contains two different alignment links for each English word.

For the specific domain, we use SF_1 and SF_2 to represent the word alignment sets in the two directions. The symbols SF, PF and MF represents the intersection set, union set and the subtraction set, respectively.

3.2 Translation Dictionary Acquisition

When we train the statistical word alignment model with the large-scale bilingual corpus in the general domain, we can get two word alignment results for the training data. By taking the intersection of the two word alignment results, we build a new alignment set. The alignment links in this intersection set are extended by iteratively adding word alignment links into it as described in [10].

Based on the extended alignment links, we build an English to Chinese translation dictionary D_1 with translation probabilities. In order to filter some noise caused by the error alignment links, we only retain those translation pairs whose translation probabilities are above a threshold δ_1 or co-occurring frequencies are above a threshold δ_2.

When we train the IBM statistical word alignment model with the small-scale bilingual corpus in the specific domain, we build another translation dictionary D_2 with the same method as for the dictionary D_1. But we adopt a different filtering strategy for the translation dictionary D_2. We use log-likelihood ratio to estimate the association strength of each translation pair because Dunning [5] proved that log-likelihood ratio performed very well on small-scale data. Thus, we get the translation dictionary D_2 by keeping those entries whose log-likelihood ratio scores are greater than a threshold δ_3.

The corpus used to build D_1 is the 320,000 sentence pairs in the general domain. The corpus used to build D_2 is the 366 sentence pairs on the manual for a medical system. By setting thresholds $\delta_1 = 0.1$, $\delta_2 = 5$ and $\delta_3 = 50$, we get two translation dictionaries, the statistics information of which is showed in Table 2.[4]

Table 2. Translation Dictionary Statistics

	D_1	D_2
Unique English Words	57,380	728
Multi-Words	18,870	28
Average Chinese Translations	2.1	1.1

[4] The thresholds are obtained to ensure the best compromise of alignment precision and recall on the testing set.

In the translation dictionary D_1, the multi-words accounts for 32.89% of the total words. In the translation dictionary D_2, the number of multi-words is small because the training data are very limited.

3.3 Word Alignment Improvement

With the statistical word alignment models and the translation dictionaries trained on the corpora in the general domain and the specific domain, we describe the algorithm to improve the domain-specific word alignment in this section.

Based on the bi-directional word alignment, we define SI as $SI = SG \cap SF$ and UG as $UG = PG \cup PF - 4 * SI$. The word alignment links in the set SI are very reliable. Thus, we directly accept them as correct links and add them into the final alignment set WA. In the set UG, there are two to four different alignment links for each word. We first examine the dictionary D_1 and then D_2 to see whether there is at least one alignment link of this word included in these two dictionaries. If it is successful, we add the link with the largest probability or the largest log-likelihood ratio score to the final set WA. Otherwise, we use two heuristic rules to select alignment links. The detailed algorithm is described in Figure 2.

Input: Alignment sets SI and UG
(1) For alignment links in SI, we directly add them into WA. (2) For each English word *i*, we first find its alignment links in UG, and then do the following: a) If there are alignment links found in the translation dictionary D_1, we add the link with the largest probability to WA. b) Otherwise, if there are alignment links found in the translation dictionary D_2, we add the link with the largest log-likelihood ratio score to WA. c) If both a) and b) fail, but three links select the same target words for the English word *i*, we add this link to WA. d) Otherwise, if there are two different kinds of links for this word: one target is a single word, and the other target is a multi-word unit and the words in the multi-word unit have no link in WA, add this multi-word alignment link to WA.
Output: Updated alignment set WA

Fig. 2. Word Alignment Adaptation Algorithm

Figure 3 lists four examples for word alignment adaptation. In example (1), the phrase "based on" has two different alignment links: one is (based on, 基于) and the other is (based, 基于). And in the translation dictionary D_1, the phrase "based on" can be translated into "基于". Thus, the link (based on, 基于) is finally selected according to rule a) in Figure 2. In the same way, the link (contrast, 造影) in example

(2) is selected with the translation dictionary D_2. The link (reconstructed, 再现) in Example (3) is obtained because there are three alignment links selecting it. For the English word "x-ray" in Example (4), we have two different links in UG. One is (x-ray, X) and the other is (x-ray, X 射线). And the single Chinese words "射" and "线" have no alignment links in the set WA. According to the rule d), we select the link (x-ray, X 射线).

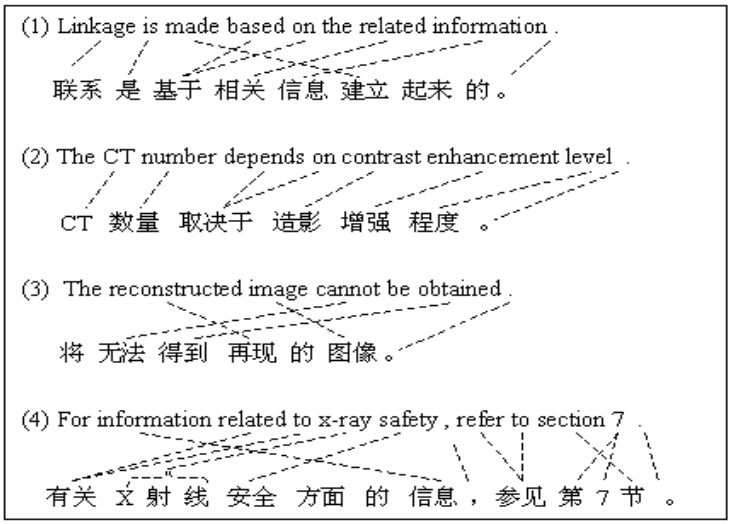

Fig. 3. Alignment Adaptation Example

4 Evaluation

In this section, we compare our methods with three other methods. The first method "Gen+Spec" directly combines the corpus in the general domain and in the specific domain as training data. The second method "Gen" only uses the corpus in the general domain as training data. The third method "Spec" only uses the domain-specific corpus as training data. With these training data, the three methods can get their own translation dictionaries. However, each of them can only get one translation dictionary. Thus, only one of the two steps a) and b) in Figure 2 can be applied to these methods. All of these three methods first get bi-directional statistical word alignment using the GIZA++ tool, and then use the trained translation dictionary to improve the statistical word alignment results. The difference between these three methods and our method is that, for each source word, our method provides four candidate alignment links while the other three methods only provides two candidate alignment links. Thus, the steps c) and d) in Figure 2 cannot be applied to these three methods.

The training data and the testing data are the same as described in Section 2.1. With the evaluation metrics described in section 2.2, we get the alignment results

shown in Table 3. From the results, it can be seen that our approach performs the best among others. Our method achieves a 21.96% relative error rate reduction as compared with the method "Gen+Spec". In addition, by comparing the results in Table 3 and those in Table 1 in Section 2.2, we can see that the precision of word alignment links is improved by using the translation dictionaries. Thus, introducing translation dictionary results in alignment precision improving while combining the alignment results of "Gen" and "Spec" results in alignment recall improving.

Table 3. Word Alignment Adaptation Results

Method	Precision	Recall	AER
Ours	0.8363	0.7673	0.1997
Gen+Spec	0.8276	0.6758	0.2559
Gen	0.8668	0.6428	0.2618
Spec	0.8178	0.4769	0.3974

Table 4. Multi-Word Alignment Results

Method	Precision	Recall	AER
Ours	0.5665	0.4083	0.5254
Gen+Spec	0.4339	0.096	0.8430
Gen	0.5882	0.083	0.8541
Spec	0.5854	0.100	0.8292

In the testing set, there are 240 multi-word alignment links. Most of the links consist of domain-specific words. Table 4 shows the results for multi-word alignment. Our method achieves much higher recall than the other three methods and achieves comparable precision. This indicates that combining the alignment results created by the "Gen" method and the "Spec" method increases the possibility of obtaining multi-word alignment links. From the table, it can be also seen that the "Spec" method performs better than both the "Gen" method and the "Gen+Spec" method on the multi-word alignment. This indicates that the "Spec" method can catch domain-specific alignment links even when trained on the small-scale corpus. It also indicates that by adding the domain-specific data into the general training data, the method "Gen+Spec" cannot catch the domain-specific alignment links.

5 Conclusion and Future Work

This paper proposes an approach to improve domain-specific word alignment through alignment adaptation. Our contribution is that, given a large-scale general bilingual corpus and a small-scale domain-specific corpus, our approach improves the domain-specific word alignment results in terms of both precision and recall. In addition, with the training data, two translation dictionaries are built to select or modify the word alignment links and to further improve the alignment results. Experimental results indicate that our approach achieves a precision of 83.63% and a recall of 76.73% for word alignment on the manual of a medical system, resulting in a relative error rate

reduction of 21.96%. This indicates that our method significantly outperforms the method only combining the general bilingual corpus and the domain-specific bilingual corpus as training data.

Our future work includes two aspects. First, we will seek other adaptation methods to further improve the domain-specific word alignment results. Second, we will also use the alignment results to build terminological lexicons and to improve translation quality and efficiency in machine translation systems.

References

1. Ahrenberg, L., Merkel, M., Andersson, M.: A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Tests. In Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th Int. Conf. on Computational Linguistics (ACL/COLING-1998) 29-35
2. Ahrenberg, L., Merkel, M., Hein, A.S., Tiedemann, J.: Evaluation of Word Alignment Systems. In Proc. of the Second Int. Conf. on Linguistic Resources and Evaluation (LREC-2000) 1255-1261
3. Brown, P.F., Della Pietra, S., Della Pietra, V., Mercer, R.: The Mathematics of Statistical Machine Translation: Parameter estimation. Computational Linguistics (1993), Vol. 19, No. 2, 263-311
4. Cherry, C., Lin, D.K.: A Probability Model to Improve Word Alignment. In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003) 88-95
5. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics (1993), Vol. 19, No. 1, 61-74
6. Ker, S.J., Chang, J.S.: A Class-based Approach to Word Alignment. Computational Linguistics (1997), Vol. 23, No. 2, 313-343
7. Langlais, P.: Improving a General-Purpose Statistical Translation Engine by Terminological Lexicons. In Proc. of the 2nd Int. Workshop on Computational Terminology (COMPUTERM-2002) 1-7
8. Melamed, D.: Automatic Construction of Clean Broad-Coverage Translation Lexicons. In Proc. of the 2nd Conf. of the Association for Machine Translation in the Americas (AMTA-1996) 125-134
9. Menezes, A., Richardson, S.D.: A Best-First Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora. In Proc. of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation (2001) 39-46
10. Och, F.J., Ney, H.: Improved Statistical Alignment Models. In Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000) 440-447
11. Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating Collocations for Bilingual Lexicons: a Statistical Approach. Computational Linguistics (1996), Vol. 22, No. 1, 1-38
12. Simard, M., Langlais, P.: Sub-sentential Exploitation of Translation Memories. In Proc. of MT Summit VIII (2001) 335-339
13. Somers, H.: Review Article: Example-Based Machine Translation. Machine Translation (1999), Vol. 14, No. 2, 113-157
14. Tufis, D., Barbu, A.M.: Lexical Token Alignment: Experiments, Results and Application. In Proc. of the Third Int. Conf. on Language Resources and Evaluation (LREC-2002) 458-465
15. Wu, D.K.: Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. Computational Linguistics (1997), Vol. 23, No. 3, 377-403