Improving Domain-Specific Word Alignment for Computer Assisted Translation

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, China, 100738
{wuhua, wanghaifeng}@rdc.toshiba.com.cn

Abstract

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. And the alignment results are applied in a computer assisted translation system to improve human translation efficiency.

1 Introduction

Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) (Brown et al., 1993). In previous alignment methods, some researchers modeled the alignments with different statistical models (Wu, 1997; Och and Ney, 2000; Cherry and Lin, 2003). Some researchers use similarity and association measures to build alignment links (Ahrenberg et al., 1998; Tufis and Barbu, 2002). 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 (Ker and Chang, 1997). 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.

This paper addresses the problem of word alignment in a specific domain, where only a small domain-specific 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 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.

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 the large-scale bilingual corpus in the general domain; (2) We train another word alignment model using the 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 method improves domain-specific word alignment in terms of both precision and recall, achieving a 21.96% relative error rate reduction.

The acquired alignment results are used in a generalized translation memory system (GTMS, a kind of computer assisted translation systems) (Simard and Langlais, 2001). This kind of system facilitates the re-use of existing translation pairs to translate documents. When translating a new sentence, the system tries to provide the pre-translated examples matched with the input and recommends a translation to the human translator, and then the translator edits the suggestion to get a final translation. The conventional TMS can only recommend translation examples on the sentential level while GTMS can work on both sentential and sub-sentential levels by using word alignment results. These GTMS are usually employed to translate various documents such as user manuals, computer operation guides, and mechanical operation manuals.

2 Word Alignment Adaptation

2.1 Bi-directional Word Alignment

In statistical translation models (Brown et al., 1993), 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 (Och and Ney, 2000). The GIZA++ toolkit [1] 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.

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

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

Where, $a_k (k = i, j)$ is the position of the source word aligned to the target word in position *k*. The set $A_k (k = i, j)$ indicates the words aligned to the same source word *k*. For example, if a Chinese word in position *j* is connect to an English word in position *i*, then $a_j = i$. And if a Chinese word in position *j* is connect to English words in position *i* and *k*, then $A_j = \{i, k\}$.

Based on the above two alignment sets, we obtain their intersection set, union set [2] and subtraction set.

Intersection: $SG = SG_1 \cap SG_2$

Union: $PG = SG_1 \cup SG_2$

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

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.

2.2 Translation Dictionary Acquisition

When we train the statistical word alignment model with a 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 (Och and Ney, 2000).

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 a limited 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 (1993) 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.

2.3 Word Alignment Adaptation Algorithm

Based on the bi-directional word alignment, we define SI as $SI = SG \cap SF$ and UG as $UG = PG \cup PF - SI$. 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.

Input: Alignment set SI and UG

(1) For alignment links in SI, we directly add them into the final alignment set WA.

(2) For each English word *i* in the UG, we first find its different alignment links, and then do the following:

a) If there are alignment links found in dictionary D_1, add the link with the largest probability to WA.

b) Otherwise, if there are alignment links found in dictionary D_2, 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 into WA.

d) Otherwise, if there are two different 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

Figure 1. Word Alignment Adaptation Algorithm

[1] It is located at http://www.isi.edu/~och/GIZA++.html

[2] 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}.

For each source word in the set UG, there are two to four different alignment links. We first use translation dictionaries to select one link among them. We first examine the dictionary D_1 and then D_2 to see whether there is at least an 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 word alignment links. The detailed algorithm is described in Figure 1.

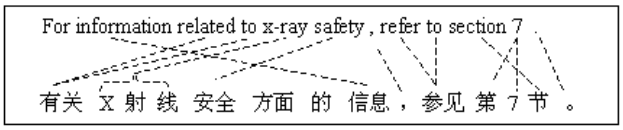

Figure 2. Alignment Example

Figure 2 shows an alignment result obtained with the word alignment adaptation algorithm. For example, for the English word "x-ray", 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).

3 Evaluation

We compare our method 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 1 can be applied to these methods. The difference between these three methods and our method is that, for each word, our method has four candidate alignment links while the other three methods only has two candidate alignment links. Thus, the steps c) and d) in Figure 1 should not be applied to these three methods.

3.1 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 medical system manual), 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.

The Chinese sentences in both the training set and the testing set are automatically segmented into words. 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.

3.2 Overall Performance

We use evaluation metrics similar to those in (Och and Ney, 2000). However, we do not classify alignment links into sure links and possible links. We consider each alignment as a sure link. 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 (Och and Ney, 2000), AER can be calculated with Equation (4). Thus, the higher the f-measure is, the lower the alignment error rate is. Thus, we will only give precision, recall and AER values in the experimental results.

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

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

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

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

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 1. Word Alignment Adaptation Results

We get the alignment results shown in Table 1 by setting the translation probability threshold to $\delta_1 = 0.1$, the co-occurring frequency threshold to $\delta_2 = 5$ and log-likelihood ratio score to $\delta_3 = 50$. From the results, it can be seen that our approach performs the best among others, achieving much higher recall and comparable precision. It also achieves a 21.96% relative error rate reduction compared to the method "Gen+Spec". This indicates that separately modeling the general words and domain-specific words can effectively improve the word alignment in a specific domain.

4 Computer Assisted Translation System

A direct application of the word alignment result to the GTMS is to get translations for sub-sequences in the input sentence using the pre-translated examples. For each sentence, there are many sub-sequences. GTMS tries to find translation examples that match the longest sub-sequences so as to cover as much of the input sentence as possible without overlapping. Figure 3 shows a sentence translated on the sub-sentential level. The three panels display the input sentence, the example translations and the translation suggestion provided by the system, respectively. The input sentence is segmented to three parts. For each part, the GTMS finds one example to get a translation fragment according to the word alignment result. By combining the three translation fragments, the GTMS produces a correct translation suggestion "系统被认为有 CT 扫描机。" Without the word alignment information, the conventional TMS cannot find translations for the input sentence because there are no examples closely matched with it. Thus, word alignment information can improve the translation accuracy of the GTMS, which in turn reduces editing time of the translators and improves translation efficiency.

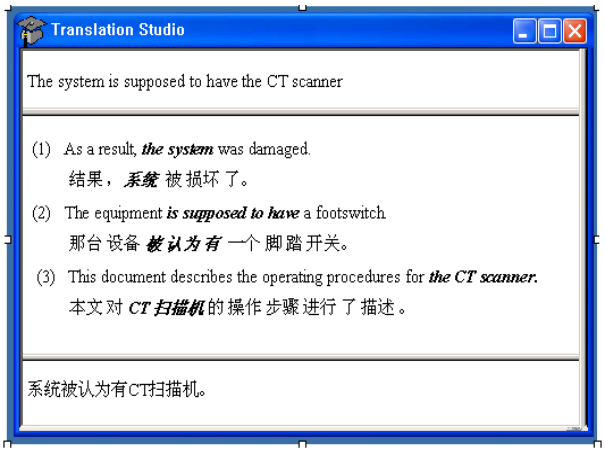

Figure 3. A Snapshot of the Translation System

5 Conclusion

This paper proposes an approach to improve domain-specific word alignment through alignment adaptation. Our contribution is that our approach improves domain-specific word alignment by adapting word alignment information from the general domain to the specific domain. Our approach achieves it by training two alignment models with a large-scale general bilingual corpus and a small-scale domain-specific corpus. Moreover, with the training data, two translation dictionaries are built to select or modify the word alignment links and 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 a user manual of a medical system, resulting in a relative error rate reduction of 21.96%. Furthermore, the alignment results are applied to a computer assisted translation system to improve translation efficiency.

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 use the alignment adaptation results in other applications.

References

Lars Ahrenberg, Magnus Merkel and Mikael Andersson. 1998. *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 International Conference on Computational Linguistics, pages 29-35.

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

Colin Cherry and Dekang Lin. 2003. *A Probability Model to Improve Word Alignment.* In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics, pages 88-95.

Ted Dunning. 1993. *Accurate Methods for the Statistics of Surprise and Coincidence.* Computational Linguistics, 19(1): 61-74.

Sue J. Ker, Jason S. Chang. 1997. *A Class-based Approach to Word Alignment*. Computational Linguistics, 23(2): 313-343.

Franz Josef Och and Hermann Ney. 2000. *Improved Statistical Alignment Models.* In Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pages 440-447.

Michel Simard and Philippe Langlais. 2001. *Sub-sentential Exploitation of Translation Memories.* In Proc. of MT Summit VIII, pages 335-339.

Dan Tufis and Ana Maria Barbu. 2002. *Lexical Token Alignment: Experiments, Results and Application.* In Proc. of the Third International Conference on Language Resources and Evaluation, pages 458-465.

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