Improving Translation Models by Applying Asymmetric Learning

Setsuo Yamada[†], **Masaaki Nagata**[†] **and Kenji Yamada**[‡]

† NTT Cyber Space Laboratories,
NTT Corporation
1-1 Hikari-no-oka, Yokosuka-shi,
Kanagawa, 239-0847 JAPAN
{yamada.setsuo, nagata.masaaki}
@lab.ntt.co.jp

‡ University of Southern California
4676 Admiralty Way, Suite 1001 Marina
del Rey, CA 90292
kyamada@isi.edu

Abstract

The statistical Machine Translation Model has two components: a language model and a translation model. This paper describes how to improve the quality of the translation model by using the common word pairs extracted by two asymmetric learning approaches. One set of word pairs is extracted by Viterbi alignment using a translation model, the other set is extracted by Viterbi alignment using another translation model created by reversing the languages. The common word pairs are extracted as the same word pairs in the two sets of word pairs. We conducted experiments using English and Japanese. Our method improves the quality of a original translation model by 5.7%. The experiments also show that the proposed learning method improves the word alignment quality independent of the training domain and the translation model. Moreover, we show that common word pairs are almost as useful as regular dictionary entries for training purposes.

1 Introduction

The statistical Machine Translation model was proposed by (Brown et al., 1988). This model has two components: a translation model(TM) and a language model. However, since the TM in (Brown et al., 1988) is based on a string-to-string noisy channel model, it is not powerful enough to treat languages that have quite different structures, such as English and Japanese. To solve this problem, several methods (Yamada and Knight, 2001; Watanabe et al., 2002) that use structural information have been proposed. However, their TMs are still not strong enough.

One simple approach to improving the quality is to add dictionary entries to training data. However, it is rare for a dictionary to include all word pairs appearing the many training sentences available. One interesting approach is seen in the TMs described in (Vogel et al., 1996; Yamada and Knight, 2001; Marcu, 2001; Watanabe et al., 2002); they set correspondences from one language to the other. Accordingly, we paid attention to both training languages. If the training languages are A and B, both $P(A|B)$ and $P(B|A)$ can be used to enhance TM performance. Although $P(A|B)$ is not the same as $P(B|A)$, we can expect that the word pairs yielded by each TM are the same.

Och et al. proposed a method that uses alignment templates from both TMs (Och et al., 1999). They improved alignment quality by combining the two alignments using a heuristic. They reported that their approach improved the translation results. However, no information was provided on the TM improvement possible by using only $P(A|B)$ and $P(B|A)$. This paper describes several experiments conducted to elucidate this effect.

In this paper, we call the word pairs extracted by asymmetrical learning common word pairs. Given that one input string in language A is $S_A = a_1...a_i...a_m$, the other input string in language B is $S_B = b_1...b_j...b_n$. It follows that word pairs extracted by alignment using $P(A|B)$ are

$$W_{ba} = \{(b_j, a_i)|i = 1..m, j = 1..n\}$$

[‡] Current affiliation is Xerox Research Centre Europe.

and the word pairs extracted by alignment using $P(B|A)$ are

$$W_{ab} = \{(a_i, b_j)|i = 1..m, j = 1..n\}.$$

Let one set of all word pairs for language A be W_A and the corresponding set of all word pairs for language B be W_B, the common word pairs are then defined as

$$W_{ab} \cap W_{ba} \text{ where } (b_j, a_i) = (a_i, b_j) \text{ in } W_A \text{ and } W_B.$$

Our approach to using common word pairs is similar to co-training methods (Yarowsky, 1995; Blum and Mitchell, 1998; Collins and Singer, 1999). This is because our method can be considered as using the common word pairs as seeds in co-training, and that common word pairs can be extracted by asymmetrical learning. Both translation model $P(A|B)$ and $P(B|A)$ are improved by the use of common word pairs.

Although we can easily apply our method to many of the TMs that have asymmetrical learning between source and target languages, this paper examines the TM in (Yamada and Knight, 2001) in order to challenge the method with languages whose structures are quite different, i.e. English and Japanese. Performance was evaluated in terms of the f-measure of word alignment quality.

The experiments indicate that our approach does improve word-level TM quality regardless of the domain and TM, its performance does, however, depend on the amount of training data. Moreover, by using only common word pairs, our approach can improve word level TM quality as if we used entries and their translations in a bilingual dictionary.

The next section explains our method in detail. Section 3 describes the experiments and the results gained. Section 4 discusses the results and provides an expansion of our approach; we conclude the paper in Section 5.

2 Improved Translation Model

This section describes the proposed learning method, which can be applied to all TMs that have asymmetric learning between source and target languages. Figure 1 shows the learning components while Figure 2 outlines the proposed learning process.

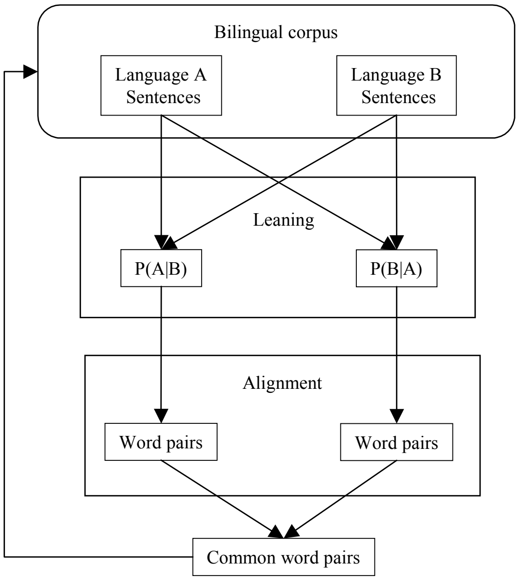

Figure 1: Learning component

1. Input language A and language B sentences from bilingual corpus.
2. Construct $P(A|B)$ and $P(B|A)$ independently.
3. Generate the most probable alignment of the training corpus for each model.
4. Extract common word pairs from both alignment results.
5. Add common word pairs to training corpus.

Figure 2: Learning process

We explain our method by applying it to the TM described in (Yamada and Knight, 2001); the input languages are English and Japanese.

In common with the regular TM learning approach, we start by preparing training sentences. We subject all input sentences to morphological and syntactical analysis before step2 because the TM in (Yamada and Knight, 2001) needs word unit and structural information. Step2 trains translation models $P(J|E)$ and $P(E|J)$ independently. These steps are common to the existing approach.

To simplify the explanation, we focus on the Japanese sentence “kare no ie ha doko ka kono atari

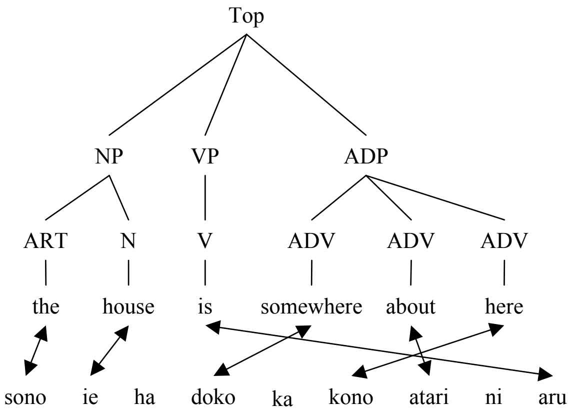

Figure 3: Example of correct alignment for E-J

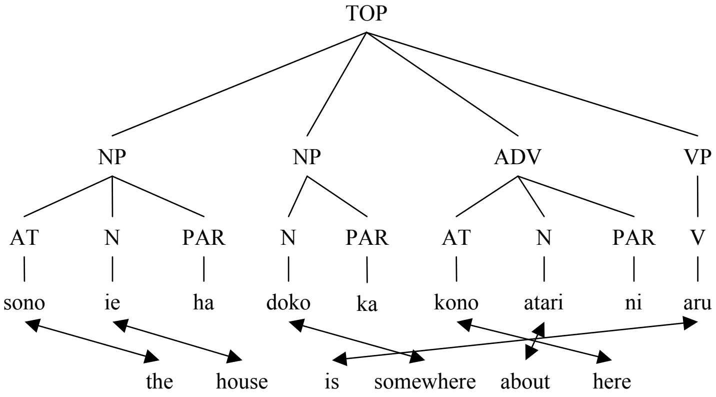

Figure 4: Example of correct alignment for J-E

ni aru" and the English sentence "His house is somewhere about here". Figure 3 shows the correct alignment for E-J, while Figure 4 shows the correct alignment for J-E. In this paper, "E-J" denotes that the input is an English parse tree and Japanese sentence; "J-E" denotes that the input is a Japanese parse tree and English sentence. [1]

Viterbi alignment is performed using $P(J|E)$ and $P(E|J)$ independently at step3. Figure 5 shows the initial E-J alignment result, Figure 6 shows the initial J-E alignment result as one example.

Step4 extracts common word pairs from the alignment results. In this example, common word pairs such as ("the", "sono"), ("house", "ie") are extracted. Since the correct word alignment is shown in figure 3 and 4, the initial word alignments are not accurate.

Step5 adds the common word pairs to the training corpus, and $P(J|E)$ and $P(E|J)$ are trained again [2] using training sentences and the common word pairs. Adding the common word pair ("here","kono"), found in another training sentence, yields the second alignment results shown in Figure 7 and Figure 8. The word alignment quality is improved. Several bad word alignments are changed into good word alignments as shown by the circle in the figures. Therefore, common word pairs can improve TM quality.

Although our basic idea is not progressive approach, our method is simpler than another method as is described in (Och et al., 1999). It is easier to apply our method to many TMs, and quality of the TMs can be improved. This is shown by experiments mentioned in the next section.

[1] In other word, "E-J" means that target language is English and source language is Japanese. ("J-E" is the reverse case.)

[2] The trained TMs are not used at the next training. That is new TMs are trained from only the bilingual corpus and common word pairs.

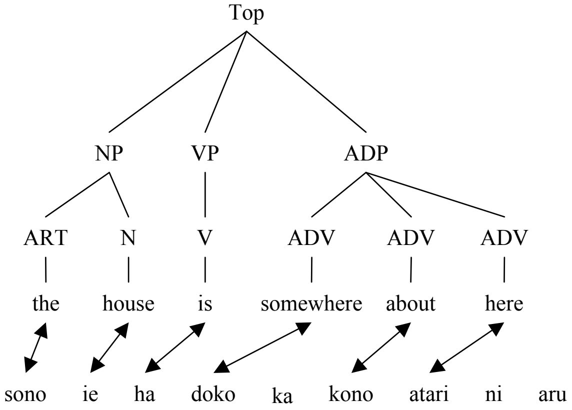

Figure 5: Initial E-J alignment

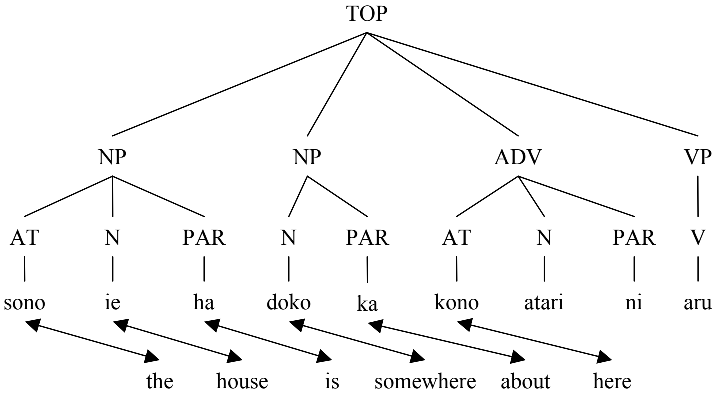

Figure 6: Initial J-E alignment

3 Experiments

This section describes the experiments conducted to quantify the improvement in word alignment quality. We selected 7,453 bilingual training sentences [3] from each of three different domains: a bilingual dictionary, scientific articles, and a newspaper. We selected 300 sentences from each domain as reference data, and aligned them by hand. The evaluation criterion was the f-measure computed from the precision and recall rates. Precision is defined as the number of correct word pairs extracted by Viterbi alignment divided by the number of word pairs extracted by Viterbi alignment. Recall is defined as the number of correct word pairs extracted by Viterbi alignment divided by the number of word pairs in the reference data.

Although our method can be applied to many TMs, in these experiments we employed the TM described in (Yamada and Knight, 2001). Because Japanese structure is quite different from English structure.

We used the English part of speech tagger described in (Brill, 2000), the English parser in (Collins, 1999) for English, and the morphological analyzer in (Asahara and Matsumoto, 2000), the parser in (Kudo and Matsumoto, 2002) for Japanese analysis. Additionally, we changed the English parse tree as in (Yamada and Knight, 2001) and slightly flattened the Japanese parse tree.

The results of the experiment indicate that our approach can improve word level TM quality regardless of the domain and TM, but its effectiveness depends on the amount of training data. Moreover, we can say that common word pairs are as useful as dictionary entries.

[3] This number equals the smallest number of sentences in the three corpora considered.

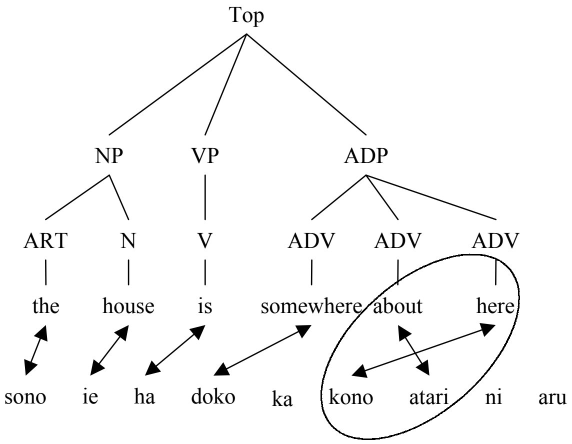

Figure 7: Second E-J alignment

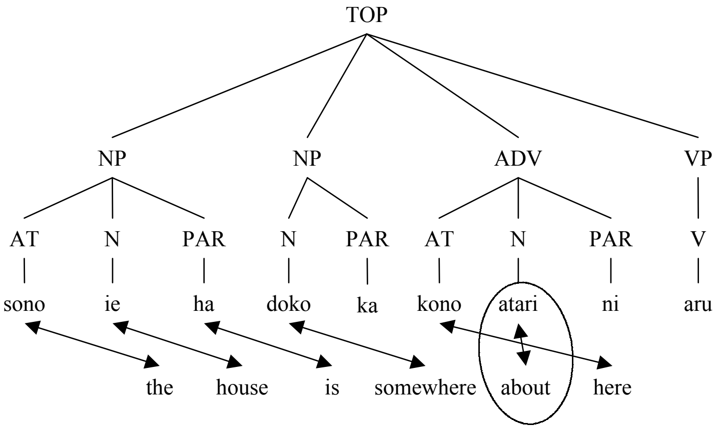

Figure 8: Second J-E alignment

3.1 The individual domains

This section shows that the proposed method can improve word alignment quality by a few percent for each of the domains examined. The results of this experiment are shown in Table 1.

“base” in the table represents the word alignment quality gained by just the original TMs $P(J|E)$ and $P(E|J)$. “new” in the table represents word alignment quality achieved by $P(J|E)$ and $P(E|J)$ after retraining by adding common word pairs to the original training sentences.

The table shows that the proposed learning method improves word alignment quality regardless of the domain and TM. However, the improvement is only 1% or so. Accordingly, we calculated paired t-statistics.

We calculated paired t-statistics for all experiments. As one example, the paired t-statistics were 0.11 for E-J and 0.03 for J-E as calculated from the result of the experiment on the dictionary sentences (“Dic.” line in Table 1). Since 0.11 exceeds 0.05, we conducted the experiment again using about 10,000 training sentences from the dictionary. The resulting paired t-statistics were 0.002 for J-E and 0.04 for E-J. This indicates that the level of significance was statistically satisfied. Results that satisfy the level of significance statistically (i.e. paired t-statistics < 0.05) are shown in bold face for all experiments. This suggests that more training sentences are needed to achieve higher statistically significant levels, especially when the initial word alignment quality is high.

3.2 Experiment on Mixed Domains

This section shows that word alignment quality can be further improved by mixing the three domains.

	E-J		J-E	
	base (%)	new (%)	base (%)	new (%)
Dic.	65.0	66.1	70.0	**71.2**
Sci.	43.0	**45.9**	38.9	**43.3**
News.	28.9	**32.7**	29.2	**33.5**

Table 1: Alignment quality for each domain

The results of this experiment are shown in Table 2.

Compared to the results in section 3.1, the dictionary domain has slightly worse quality after mixing. The reason is that the original dictionary sentence pairs have high sentence level alignment, so using the results from the other domains degrades the alignment.

These tables also show that the quality of the word alignment yielded by the proposed learning method improves with the number of training sentences.

	E-J		J-E	
	base (%)	new (%)	base (%)	new (%)
Dic.	64.3	**66.4**	66.2	**69.6**
Sci.	46.2	**48.1**	42.9	**48.6**
News.	32.7	**37.1**	31.1	**35.0**

Table 2: Alignment quality with domain mixing

3.3 Experiment on Word Pairs

This section compares the performance possible with just common word pairs to that achieved with dictionary entries. We employed a different dictionary in this experiment as the source of the dictionary entries.

The results of this experiment are shown in Table 3. "Entries" in the table means the case of applying only entries in the dictionary and "Common Pairs" means the case of applying only common word pairs extracted from dictionary example sentences.

The table indicates that the performances are almost the same. Therefore, we can say that the performance achieved by common word pair is similar that possible with dictionary entries. That is, if we do not have a dictionary for the field of interest, using sentence pairs can improve the word alignment quality to almost the same extent as would the dictionary entries if they were available.

	E-J (%)	J-E (%)
Entries	65.7	70.4
Common pairs	66.1	71.2

Table 3: Alignment quality when using dictionary entries

3.4 Iterative Applying Common Word Pairs

We iterated the leaning process in Figure 2. In this experiment we added new common word pairs extracted at each iteration to the original bilingual corpus, because we expected that some inappropriate common word pairs extracted by the previous process were removed. Therefore the set of common word pairs extracted in the first iteration is not a subset of common word pairs extracted in second iteration.

As shown in Figure 9, the quality basically saturates after one or two iterations. This means that our approach does not need to iterate the learning process many times.

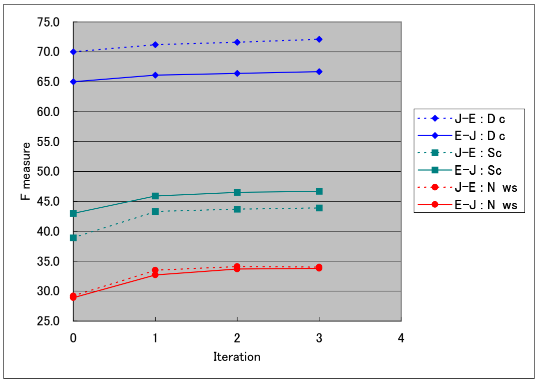

Figure 9: Training iteration experiment

4 Discussion

It is easy to apply our approach to the many TMs that have asymmetric learning between source and target languages. We note, however, that the second alignment results described in section 2 are still not perfect. One problem is that the Japanese "aru" is far from the English "is". To make the improvement more effective, we might need to directly utilize TM probability as is described in (Och et al., 1999). This is because the TM has many useless word pairs. For instance, it is possible to extend the proposed approach by weighting each common word pair with a probability according to the part of speech involved.

The above text discussed only word alignment, but we can also imagine expanding our approach to cover phrase alignment or n-to-n word alignment. Phrase alignment or n-to-n word alignment approaches for statistical translation (e.g. (Och et al., 1999; Marcu, 2001; Varea et al., 2002; Watanabe et al., 2002)) have been proposed. Adding the proposed approach to these algorithms would improve their performance.

5 Conclusion

We have proposed a method that improves translation model quality by using common word pairs extracted by asymmetric learning. If $P(A|B)$ and $P(B|A)$ are training models (TMs), we can get common word pairs from both results of Viterbi alignment using $P(A|B)$ and $P(B|A)$. By retraining $P(A|B)$ and $P(B|A)$ using the original training sentences plus the above common word pairs, the quality of both $P(A|B)$ and $P(B|A)$ are improved.

We conducted experiments on English and Japanese material by applying the proposed method to the TM proposed in (Yamada and Knight, 2001) for three different domains: a bilingual dictionary, scientific articles, and a newspaper. Experiments were evaluated by the f-measure of word alignment quality. The results showed that the proposed method improves word alignment by 5.7%.

It is easy to apply our approach to other TMs. We can say that our approach improves word level TM quality regardless of the domain and TM; its performance does, however, depend on the amount of training data used. Moreover, by using only common word pairs, the proposed approach can improve word level TM quality to the same extent as is possible if word pairs from a bilingual dictionary are used.

We intend to expand and refine our approach. For example, we intend to support phrase level alignment or n-to-n word alignment and apply it to another TM.

References

Masayuki Asahara and Yuji Matsumoto. 2000. Extended models and tools for high-performance part-of-speech tagger. In *Proc. of Coling 2000*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proc. of COLT-98*, pages 92–100.

Eric Brill. 2000. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(2).

Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, Robert L. Mercer, and Paul S. Roossin. 1988. A statistical approach to language translation. In *Proc. of Coling-88*.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proc. of EMNLP/WVLC-99*, pages 100–110.

Michael Collins. 1999. Head-driven statistical models for natural language parsing. *Ph.D thesis*.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proc. of Coling 2002*.

Daniel Marcu. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proc. of ACL 2001*, pages 386–393.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of EMNLP/WVLC-99*, pages 20–28.

Ismael García Varea, Franz J. Och, and Francisco Casacuberta. 2002. Improving alignment quality in statistical machine translation using context-dependent maximum entropy models. In *Proc. of Coling 2002*, pages 1051–1054.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proc. of ACL-96*, pages 836–841.

Taro Watanabe, Kenji Imamura, and Eiichiro Sumita. 2002. Statistical machine translation based on hierarchical phrase alignment. In *Proc. of TMI 2002*, pages 188–198.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proc. of ACL 2001*, pages 523–530.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivailing supervised methods. In *Proc. of ACL-95*, pages 189–196.