

# Using Comparable Corpora to Adapt a Translation Model to Domains

**Hiroyuki Kaji, Takashi Tsunakawa, Daisuke Okada**

Department of Computer Science, Shizuoka University  
3-5-1 Johoku, Naka-ku, Hamamatsu-shi, 432-8011, Japan  
{kaji, tuna}@inf.shizuoka.ac.jp

## Abstract

Statistical machine translation (SMT) requires a large parallel corpus, which is available only for restricted language pairs and domains. To expand the language pairs and domains to which SMT is applicable, we created a method for estimating translation pseudo-probabilities from bilingual comparable corpora. The essence of our method is to calculate pairwise correlations between the words associated with a source-language word, presently restricted to a noun, and its translations; word translation pseudo-probabilities are calculated based on the assumption that the more associated words a translation is correlated with, the higher its translation probability. We also describe a method we created for calculating noun-sequence translation pseudo-probabilities based on occurrence frequencies of noun sequences and constituent-word translation pseudo-probabilities. Then, we present a framework for merging the translation pseudo-probabilities estimated from in-domain comparable corpora with a translation model learned from an out-of-domain parallel corpus. Experiments using Japanese and English comparable corpora of scientific paper abstracts and a Japanese-English parallel corpus of patent abstracts showed promising results; the BLEU score was improved to some degree by incorporating the pseudo-probabilities estimated from the in-domain comparable corpora. Future work includes an optimization of the parameters and an extension to estimate translation pseudo-probabilities for verbs.

## 1. Introduction

There has been a surge in research on statistical machine translation (SMT). SMT has an advantage in that it learns a translation model from a parallel corpus in a subject domain (Brown et al., 1993). However, it suffers from the limited availability of large parallel corpora. Therefore, a method should be developed for learning a translation model from bilingual comparable corpora, which are far more available than parallel corpora for many language pairs and in many subject domains.

We describe a method for calculating translation pseudo-probabilities from a bilingual dictionary and very weakly comparable corpora, i.e., a pair of source language (SL) and target language (TL) monolingual corpora in the same domain. Because comparable corpora do not contain correspondence even between sentences, we cannot use any translation-probability estimation methods based on word-for-word alignment.

We turned our attention to word associations that suggest particular senses or translations of a polysemous word. Comparable corpora allow us to determine which word associations suggest which translations of a polysemous word (Kaji and Morimoto, 2002). Assuming that the more word associations that suggested a translation, the higher the probability of the translation would be, we created a method for estimating word-for-word translation pseudo-probabilities. We also describe a method we created for estimating phrase-for-phrase translation pseudo-probabilities because we aim at using the estimated pseudo-probabilities in phrase-based SMT (Koehn et al., 2003). Note that, in this paper, words and phrases are restricted to nouns and noun sequences.

In addition, we present a framework combining the pseudo-probabilities estimated from in-domain comparable corpora with the translation model learned from an out-of-domain parallel corpus. We describe our demonstration of the feasibility of our method through a Japanese-to-English SMT experiment using an out-of-domain parallel corpus of patent abstracts and in-domain comparable corpora of scientific paper abstracts as well as another experiment using a smaller in-domain parallel corpus and larger in-domain comparable corpora.

## 2. Estimating translation pseudo-probabilities from comparable corpora

### 2.1 Basic idea

The underlying assumptions of our method are as follows:

- (i) Translations of words associated with each other in a language are also associated with each other in another language (Rapp, 1995). For example, two English words “tank” and “soldier” are associated with each other and, at the same time, their Japanese translations “戦車[*SENSHA*]” and “兵士[*HEISHI*]” are associated with each other.
- (ii) A polysemous word exhibits only one sense per word association (Yarowsky, 1993). For example, a polysemous word “tank” exhibits the “military vehicle” sense when it is associated with “soldier,” while it exhibits the “container for liquid or gas” sense when it is associated with “gasoline.”

In addition, we assume different translations of a word indicate different senses of the word. Under these assumptions, we can determine which of the words

associated with an SL word suggests which of its translations by aligning word associations across languages with the assistance of a bilingual dictionary. For example, the alignment of an English word association (tank, soldier) with its Japanese counterpart (戦車[SENSHA], 兵士[HEISHI]) allows us to determine that the associated word “soldier” suggests the translation “戦車[SENSHA],” while the alignment of another English word association (tank, gasoline) with its Japanese counterpart (タンク[TANKU], ガソリン[GASORIN]) allows us to determine that the associated word “gasoline” suggests the translation “タンク[TANKU].”

This naive method suffers from the following two difficulties.

(a) Aligning word associations often fails because of the disparity in topical coverage between two language corpora as well as the incomplete coverage of the bilingual dictionary consulted. Suppose, for example, a word association (tank, Chechen) is obtained from an English corpus. However, its counterpart (戦車[SENSHA], チェチェン[CHECHEN]) may not be obtained from a Japanese corpus. Moreover, even when (戦車[SENSHA], チェチェン[CHECHEN]) is also obtained from the Japanese corpus, a pair of translation equivalents (Chechen, チェチェン[CHECHEN]) may not be contained in an English-Japanese dictionary. Thus, we cannot necessarily determine that the associated word “Chechen” suggests the translation “戦車[SENSHA].”

(b) Incorrect word-association alignment often happens because of incidental word-for-word correspondence between word associations that do not really correspond to each other. For example, “tank” and “troop” correspond to “水槽[SUISOU]” and “群れ[MURE],” respectively, although an English word association (tank, troop) does not correspond to a Japanese word association (水槽[SUISOU], 群れ[MURE]). Thus, the word association (tank, troop) is aligned not only with its correct counterpart (戦車[SENSHA], 隊[TAI]) but also with a spurious counterpart (水槽[SUISOU], 群れ[MURE]). As a result, we cannot determine which of the translations “戦車[SENSHA]” and “水槽[SUISOU]” the associated word “troop” suggests.

To overcome these difficulties, we focus on the fact that two words associated with a third word are likely to suggest the same sense of the third word when they are also associated with each other. For example, “troop” and “soldier,” both of which are associated with “tank” and which are also associated with each other, suggest the same translation “戦車[SENSHA]” of “tank.”

We define a correlation between an associated word and a translation using the correlations between other associated words and the translation (Kaji and Morimoto, 2002). Note that this definition is recursive; e.g., the

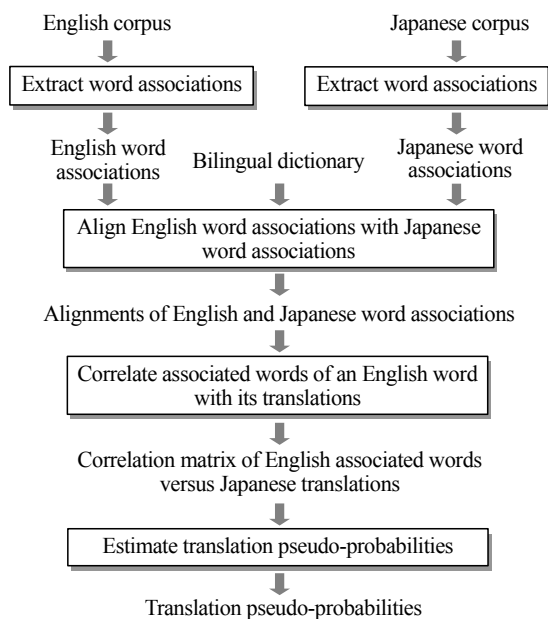


Fig. 1: Overview of our method for estimating word translation pseudo-probabilities

formula defining the correlation between “troop” and “戦車[SENSHA]” includes the correlation between “soldier” and “戦車[SENSHA],” and vice versa. According to such recursive definitions, we iteratively calculate a correlation matrix of associated words versus translations for each SL word. Finally, we determine that an associated word suggests a translation with which it has the highest correlation.

## 2.2 Calculating noun translation pseudo-probabilities

Our method consists of the following four steps (See Fig. 1):

### (1) Extract word associations

The translation of a noun is suggested by various types of associated words including topically associated ones. Therefore, we extract pairs of words co-occurring in a medium-sized window and calculate pointwise mutual information for word pairs. Pointwise mutual information  $MI(x, x')$  of words  $x$  and  $x'$  is defined as:

$$MI(x, x') = \log_2 \frac{h(x, x') / \sum_{x, x'} h(x, x')}{(g(x) / \sum_x g(x)) (g(x') / \sum_{x'} g(x'))},$$

where  $g(x)$  is the occurrence frequency of  $x$  in a corpus, and  $h(x, x')$  is the co-occurrence frequency of  $x$  and  $x'$  in the corpus. The pointwise mutual information is calculated for every pair of words whose occurrence frequencies and co-occurrence frequency are not less than their respective thresholds predetermined. Word associations are extracted by selecting word pairs with

pointwise mutual information larger than a predetermined threshold. In the experiments described in Sec. 4, the window size was set to 10 content words on either side of the current word, and the thresholds for occurrence frequency, co-occurrence frequency, and pointwise mutual information were set to 10, 6, and 0, respectively.

## (2) Align word associations across languages

An SL word association  $(f, f')$  is aligned with a TL word association  $(e, e')$  if and only if a bilingual dictionary  $D$  contains two pairs of translation equivalents,  $(f, e)$  and  $(f', e')$ . In the following,  $AL$  denotes the sets of word-association alignments obtained. That is,

$$AL = \{((f, f'), (e, e')) \mid (f, f') \in A_f, (e, e') \in A_e, (f, e) \in D, (f', e') \in D\},$$

where  $A_f$  and  $A_e$  denote the sets of word associations extracted from SL and TL corpora, respectively.

## (3) Calculate a correlation matrix of associated words vs. translations for an SL word

The correlation between the  $i$ -th word  $f'(i)$  associated with an SL word  $f$  and its  $j$ -th translation  $e(j)$  is defined as follows<sup>1</sup>:

$$C_n(f'(i), e(j)) = \frac{\sum_{\substack{(f, f'') \in A_f \\ (f'(i), f'') \in A_f}} MI(f'(i), f'') \cdot C_{n-1}(f'', e(j))}{\max_k \sum_{\substack{(f, f'') \in A_f \\ (f'(i), f'') \in A_f}} MI(f'(i), f'') \cdot C_{n-1}(f'', e(k))},$$

where suffix  $n$  indicates the iteration cycle. The correlations are calculated iteratively starting with the following initial values:

$$C_0(f'(i), e(j)) = \begin{cases} \frac{\delta(f'(i), e(j))}{\sum_k \delta(f'(i), e(k))} & \dots \sum_k \delta(f'(i), e(k)) \neq 0 \\ 0 & \dots otherwise \end{cases},$$

where  $\delta(f'(i), e(j))$  is a binary function whose values are determined according to the results of word-association alignment, i.e.,

<sup>1</sup> We modified the definition given in (Kaji and Morimoto, 2002). The original formula includes an additional term based on alignment of triplets of words associated with each other, which causes convergence of correlation values from uniform initial values. We omitted that term because it proved to be ineffective for a pair of corpora with large disparity in topical coverage. We also changed the initial values to those determined according to the results of word-association alignment so that the correlation values would converge.

| <i>plant</i> | 装置   | 設備          | 植物          | 工場   | プラント        | 苗    | 植木   |
|--------------|------|-------------|-------------|------|-------------|------|------|
| activity     | 0.02 | 0.03        | <b>2.10</b> | 0.20 | 0.03        | 0.01 | 0.02 |
| bacteria     | 0.02 | 0.03        | <b>1.98</b> | 0.01 | 0.02        | 0.27 | 0.02 |
| boiler       | 0.05 | 2.70        | 0.05        | 0.03 | <b>2.73</b> | 0.03 | 0.04 |
| coal         | 0.87 | <b>2.35</b> | 1.70        | 0.68 | 2.06        | 0.65 | 0.99 |
| computer     | 0.55 | 0.71        | 0.02        | 0.49 | <b>0.73</b> | 0.01 | 0.01 |
| control      | 0.47 | 0.51        | 0.17        | 0.15 | <b>0.62</b> | 0.06 | 0.01 |
| culture      | 0.03 | 0.05        | <b>3.26</b> | 0.23 | 0.12        | 0.77 | 0.88 |
| environment  | 0.76 | 1.25        | <b>1.32</b> | 0.03 | 0.05        | 0.23 | 0.03 |
| failure      | 0.93 | 1.22        | 0.03        | 0.53 | <b>1.43</b> | 0.01 | 0.01 |
| flower       | 0.04 | 0.06        | <b>4.02</b> | 0.04 | 0.04        | 1.23 | 1.70 |
| :            | :    | :           | :           | :    | :           | :    | :    |

Translation pseudo-probability .047 .241 .423 .022 .223 .022 .022

Fig. 2: Example correlation matrix of associated words versus translations

$$\delta(f'(i), e(j)) = \begin{cases} 1 & \dots \exists e'. ((f, f'(i)), (e(j), e')) \in AL \\ 0 & \dots otherwise \end{cases}.$$

## (4) Estimate translation pseudo-probabilities

The translation pseudo-probability of a translation is basically defined as the proportion of associated words having the highest correlations with the translation. To avoid zero-probabilities, we give every translation a small fragment  $\varepsilon$  ( $\varepsilon$  was set to 0.025 in the experiments described in Sec. 4.). Thus, the pseudo-probability of an SL word  $f$  being translated to its  $j$ -th translation  $e(j)$  is calculated as:

$$P_{ps}(e(j) | f) = \frac{|S(e(j))|/N + \varepsilon}{\sum_k (|S(e(k))|/N + \varepsilon)},$$

where  $N$  is the total number of words associated with  $f$ , and  $S(e(j))$  denotes a set consisting of all associated words having the highest correlation with  $e(j)$ , i.e.,

$$S(e(j)) = \{f'(i) \mid \forall k (\neq j). C(f'(i), e(j)) \geq C(f'(i), e(k))\}.$$

Note that the correlation matrix of associated words versus translations cannot be calculated for a low-frequency SL word. For such a word, all of its translation candidates in the bilingual dictionary are given a uniform pseudo-probability, i.e., the inverse of the number of its translation candidates.

An example correlation matrix and estimated translation pseudo-probabilities are shown in Fig. 2; the correlation matrix was calculated for the SL word “plant” from English and Japanese patent abstract corpora.

## 2.3 Calculating noun-sequence translation pseudo-probabilities

The method described in Sec. 2.2 estimates translation pseudo-probabilities for words contained in a bilingual dictionary. Because bilingual dictionaries hardly ever contain phrases consisting of two or more words, we also created a method for estimating phrase-for-phrase translation pseudo-probabilities, where phrases are limited to noun sequences. Our method consists of the following two steps.

### (1) Extract SL noun sequences and their translation candidates

All noun sequences with occurrence frequencies not less than a predetermined threshold are extracted both from the SL corpus and from the TL corpus. Then, for an SL noun sequence  $F = f_1 f_2 \dots f_m$ , a subset consisting of all TL noun sequences generated by a compositional translation saving word order, i.e.,

$$\{E(j) = e_1(j)e_2(j)\dots e_m(j) \mid (f_i, e_i(j)) \in D, i = 1, \dots, m, j = 1, \dots, n\},$$

are collected as its translation candidates.

### (2) Estimate translation pseudo-probabilities

We have two methods for estimating translation probabilities for noun sequences.

(a) Give a translation candidate a probability proportional to its occurrence frequency. That is,

$$P_1(E(j) \mid F) = \frac{g(E(j))}{\sum_{k=1}^n g(E(k))},$$

where  $g(E(j))$  is the occurrence frequency of the  $j$ -th translation candidate  $E(j)$ . Note that not all translation candidates are correct translations even if they occur in the TL corpus; a large probability may be given to an incorrect translation.

(b) Give a translation candidate a probability proportional to the product of the constituent-word translation pseudo-probabilities, under the assumption that constituent-word translation probabilities are independent. That is,

$$P_2(E(j) \mid F) = \frac{\prod_{i=1}^m P_{ps}(e_i(j) \mid f_i)}{\sum_{k=1}^n \prod_{i=1}^m P_{ps}(e_i(k) \mid f_i)},$$

where  $P_{ps}(e_i(j) \mid f_i)$  is the word translation pseudo-probability estimated by the method described in Sec. 2.2. Note that the word translation pseudo-probabilities are not well-founded for low-frequency words. In addition, the independence assumption is actually not satisfied.

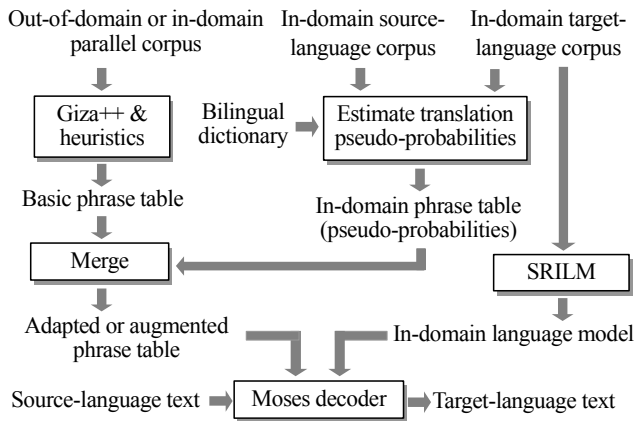


Fig. 3: Phrase-based SMT framework using translation pseudo-probabilities estimated from comparable corpora

Because both methods have their own weaknesses, we define the noun-sequence translation pseudo-probability as the geometric mean of the two probabilities. That is,

$$P_{ps}(E(j) \mid F) = \sqrt{P_1(E(j) \mid F) \cdot P_2(E(j) \mid F)}.$$

## 3. Using translation pseudo-probabilities in phrase-based SMT

We suppose the following two cases about available corpora:

- (i) Out-of-domain parallel corpus plus in-domain comparable corpora: No parallel corpora are available in the domain of texts to be translated, but both SL and TL monolingual corpora are available.
- (ii) In-domain parallel corpus plus in-domain larger comparable corpora: A parallel corpus is available in the domain, but it is not large. Much larger SL and TL monolingual corpora are available.

Figure 3 outlines a phrase-based SMT framework using the translation pseudo-probabilities estimated from comparable corpora. A basic phrase table is produced from an out-of-domain or in-domain parallel corpus by using Giza++ (Och and Ney, 2003) and the *grow-diag-final* heuristics (Koehn et al., 2003). Another phrase table is produced from in-domain SL and TL monolingual corpora and a bilingual dictionary by using the method described in the previous section. These two phrase tables are merged into one by taking a weighted mean of the two probabilities (The two tables were given the same weight in the experiment described in Sec. 4.). The in-domain TL monolingual corpus is also used to learn a TL language model with SRILM toolkits (Stolcke, 2002). The Moses decoder translates in-domain texts according to the merged phrase table as well as the TL language model.

## 4. Experiments

### 4.1 Experimental setting

We conducted Japanese-to-English translation experiments with JST’s scientific-paper abstract corpus and JAPIO’s patent abstract corpus. JST’s corpus, which consists of Japanese and English abstracts of wide-ranging comparability, from parallel to unrelated, as well as abstracts in either language, was treated as comparable corpora. In contrast, JAPIO’s corpus, which consists of Japanese patent abstracts and their English translations, was treated as a parallel corpus.

The following corpora and dictionary were compiled for two experiments: Experiment A (Out-of-domain parallel corpus plus in-domain comparable corpora) and Experiment B (In-domain parallel corpus plus in-domain larger comparable corpora).

- Training corpora
  - Out-of-domain parallel corpus (only for Experiment A): 20,000 pairs of patent abstracts in the physics (Japanese: 5.32 Mbytes, English: 4.54 Mbytes)
  - In-domain parallel corpus (only for Experiment B): 20,000 pairs of parallel sentences in the chemistry (Japanese: 3.61 Mbytes, English: 3.17 Mbytes), i.e., pairs of Japanese and English sentences having high similarity—ones extracted from the in-domain comparable corpora mentioned next<sup>2</sup>.
  - In-domain comparable corpora: Abstracts of scientific papers in the chemistry (Japanese: 151,958 abstracts (90.8 Mbytes), English: 102,730 abstracts (64.9 Mbytes)), excluding sentences extracted as the test corpus next.
- Test corpus: 1,000 pairs of parallel sentences, i.e., pairs of Japanese and English sentences having high similarity—ones extracted from in-domain comparable corpora<sup>2</sup>. Note that the Japanese sentences were used as input to the MT and the English sentences as a reference.
- Japanese-English noun dictionary: 333,656 pairs of translation equivalents between 163,247 Japanese and 93,727 English nouns, obtained by merging the EDR, EIJIRO, and EDICT Japanese-English dictionaries.

In both Experiments A and B, we evaluated our method in four cases using a different volume of in-domain comparable corpora, i.e., all or half of the Japanese corpus and all or half of the English corpus. Two baseline methods were also evaluated: a baseline without a dictionary that uses the phrase table learned from the parallel corpus, and a baseline with a dictionary that uses a mixture of the phrase table learned from the parallel corpus and a supplementary phrase table produced by giving uniform probabilities to all translation candidates

<sup>2</sup> The extraction was done automatically by using a method similar to that used by Utiyama and Isahara (2003).

Table 1: Experimental results

(a) Experiment A—Out-of-domain parallel corpus plus in-domain comparable corpora

| Method                  |                | BLEU-4 |
|-------------------------|----------------|--------|
| Our method              | J:all, E:all   | 13.30  |
|                         | J:half, E:all  | 13.19  |
|                         | J:all, E:half  | 13.21  |
|                         | J:half, E:half | 13.27  |
| Baseline w/o dictionary |                | 11.42  |
| Baseline w/ dictionary  |                | 12.94  |

(b) Experiment B—In-domain parallel corpus plus in-domain larger comparable corpus

| Method                  |                | BLEU-4 |
|-------------------------|----------------|--------|
| Our method              | J:all, E:all   | 16.82  |
|                         | J:half, E:all  | 16.70  |
|                         | J:all, E:half  | 16.78  |
|                         | J:half, E:half | 16.71  |
| Baseline w/o dictionary |                | 16.37  |
| Baseline w/ dictionary  |                | 16.32  |

contained in the Japanese-English noun dictionary. Note that the TL language model learned from the whole TL monolingual corpus was used commonly in all cases involving our method and the baseline methods.

### 4.2 Experimental results

Table 1 shows the BLEU-4 scores (Papineni et al., 2002) of our method in the four cases and the two baseline methods; the scores were calculated with one reference translation per test sentence.

In Experiment A, our method improved the score to some degree compared with the baseline w/o dictionary but improved it only a little compared with the baseline w/ dictionary; this indicates the need to improve the accuracy of a translation pseudo-probability estimate. The effect of the difference in volume of in-domain comparable corpora remains unclear; the size of comparable corpora used in the experiment might be too small.

Our method slightly improved the BLEU-4 score in Experiment B, in which the baseline was naturally high. However, the difference between our method and the baseline w/ dictionary, which performed slightly below the baseline w/o dictionary, indicates the promise of our method. We need to conduct an experiment with much larger comparable corpora.

## 5. Discussions

The experimental results demonstrated that our method is feasible, but many issues need to be solved. The directions for improvement and extension are discussed next.

### **(1) Optimization of the parameters**

Our method has many parameters, including the window size and thresholds for word occurrence frequency, co-occurrence frequency, and pointwise mutual information, all of which affect the correlation matrix of associated words vs. translations. The experiment was carried out with a parameter setting that was not necessarily optimal. How to optimize the values for the parameters remains unsolved.

### **(2) Alternatives for word-association measure**

In another study, we obtained results indicating that pointwise mutual information is not the most suitable for acquiring word associations. We need to compare it with alternatives such as log-likelihood ratio and the Dice coefficient.

### **(3) Refinement of the definition of translation pseudo-probability**

The present definition of word translation pseudo-probability is too rough. We need to consider the frequencies of associated words as well as the dependence among associated words. Also, the strategy assigning an associated word to only one translation should be reconsidered.

### **(4) Estimate of verb translation pseudo-probabilities**

The most effective clue for determining the appropriate translation of a verb is its arguments, particularly its object noun. Thus, the co-occurrence in a window needs to be replaced with syntactic co-occurrence to extract verb-noun associations from corpora. The definition of the correlation between associated words and translations also needs to be modified; instead of the heuristics described in Sec. 2.1, it should be based on other heuristics where two nouns associated with a verb are likely to suggest the same sense of the verb when they belong to the same semantic class.

Finally, we mention the language pairs to which our method is applicable. The method requires not only comparable corpora but also a wide-coverage bilingual dictionary, which seems to restrict its applicability. However, bilingual dictionaries are usually more available than parallel corpora for many language pairs. In addition, our method accepts a combination of two bilingual dictionaries via a hub language, usually English, as a bilingual dictionary; although a combination of bilingual dictionaries contains spurious translations due to ambiguous intermediary words, such translations are given small translation pseudo-probabilities (Kaji et al., 2008). Thus, our method is applicable to many domains as well as to many language pairs.

## **6. Related work**

Research on the use of comparable corpora in machine translation goes back to the mid 90s. A lot of studies on

bilingual lexicon acquisition from comparable corpora have been reported (Rapp, 1995; Kaji and Aizono, 1996; Tanaka and Iwasaki, 1996; Fung and Yee, 1998; Rapp, 1999). Most of them, based on the assumption that translations of words that co-occur in one language also co-occur in the other language, extract pairs of words that are translations of each other. However, they do not estimate translation probabilities.

To the best of our knowledge, Koehn and Knight (2000) were the first to propose estimates of translation probabilities from comparable corpora and a bilingual dictionary. Their method using an EM algorithm could produce translation probabilities greatly affected by the occurrence frequencies of translation candidates in the TL corpus. In contrast, our method produces translation pseudo-probabilities that reflect the distribution of the senses of the SL word in the SL corpus. We believe that this feature of our method is important because the texts to be translated are in source language, not in target language.

Methods for extracting a collection of parallel sentence pairs from bilingual comparable corpora have also been proposed (Zhao and Vogel, 2002; Utiyama and Isahara 2003; Fung and Cheung, 2004; Munteanu and Marcu, 2005); extracted parallel sentences are used to learn a translation model with a conventional method based on word-for-word alignment. This approach obviously is applicable only to closely comparable corpora. In contrast, our method is applicable even to a pair of unrelated monolingual corpora.

## **7. Conclusion**

We created a method for estimating translation pseudo-probabilities from a bilingual dictionary and bilingual comparable corpora. The essence of our method is to calculate pairwise correlations between associated words of a source-language word, presently restricted to a noun, and its translations. Based on the assumption that the more associated words a translation is correlated with, the higher its translation probability, the method calculates word translation pseudo-probabilities. In addition, we created a method for estimating noun-sequence translation pseudo-probabilities based on the occurrence frequencies of noun sequences and the constituent-word translation pseudo-probabilities.

We also presented a framework in which pseudo-translation probabilities estimated from in-domain comparable corpora are merged with a translation model learned from an out-of-domain parallel corpus. Experiments using Japanese and English comparable corpora of scientific paper abstracts and a Japanese-English parallel corpus of patent abstracts showed promising results; the BLEU score was improved to some degree by using the pseudo-probabilities estimated from in-domain comparable corpora. Future work includes optimizing the parameters and extending

the method to estimate translation pseudo-probabilities for verbs.

## 8. Acknowledgements

This work was supported in part by the Special Coordination Funds for Promoting Science and Technology of MEXT (the Ministry of Education, Culture, Sports, Science and Technology – Japan). We were permitted to use the scientific paper abstract corpus and the patent abstract corpus in the experiments by courtesy of JST (the Japan Science and Technology Agency) and JAPIO (the Japan Patent Information Organization), respectively.

## 9. References

- Brown, Peter F., 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.
- Fung, Pascale and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 57-63.
- Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics / the 17th International Conference on Computational Linguistics*, pp. 414-420.
- Kaji, Hiroyuki and Toshiko Aizono. 1996. Extracting word correspondences from bilingual corpora based on word co-occurrence information. *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 23-28.
- Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 411-417.
- Kaji, Hiroyuki, Shin'ichi Tamamura, and Dashtseren Erdenebat. 2008. Automatic construction of a Japanese-Chinese dictionary via English. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 699-706.
- Koehn, Philipp and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. *Proceedings of the 17th National Conference on Artificial Intelligence*, pp. 711-715.
- Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127-133.
- Munteanu, Dragos S. and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4): 477-504.
- Och, Franz J. and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1): 19-51.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318.
- Rapp, Reinhard. 1995. Identifying word translations in non-parallel texts. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.
- Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 519-526.
- Stolcke, Andrea. 2002. SRILM — An extensible language modeling toolkit. *Proceedings of International Conference on Spoken Language Processing*, pp. 901-904.
- Tanaka, Kumiko and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 580-585.
- Utiyama, Masao and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 72-79.
- Yarowsky, David. 1993. One sense per collocation. *Proceedings of ARPA Human Language Technology Workshop*, pp. 266-271.
- Zhao, Bing and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. *Proceedings of the 2002 IEEE International Conference on Data Mining*, pp. 745-748.