Journal of Information Processing Vol.23 No.3 344–352 (May 2015)

[DOI: 10.2197/ipsjjip.23.344]

**Regular Paper**

# Extraction of Potentially Useful Phrase Pairs for Statistical Machine Translation

Juan Luo[1,a)] Yves Lepage[1,b)]

Received: July 29, 2014, Accepted: January 7, 2015

***Abstract:*** Over the last decade, an increasing amount of work has been done to advance the phrase-based statistical machine translation model in which the method of extracting phrase pairs consists of word alignment and phrase extraction. In this paper, we show that, for Japanese-English and Chinese-English statistical machine translation systems, this method is indeed missing potentially useful phrase pairs which could lead to better translation scores. These potentially useful phrase pairs can be detected by looking at the segmentation traces after decoding. We choose to see the problem of extracting potentially useful phrase pairs as a two-class classification problem: among all the possible phrase pairs, distinguish the useful ones from the not-useful ones. As for any classification problem, the question is to discover the relevant features which contribute the most. Extracting potentially useful phrase pairs resulted in a statistically significant improvement of 7.65 BLEU points in English-Chinese and 7.61 BLEU points in Chinese-English experiments. A slight increase of 0.94 BLEU points and 0.4 BLEU points is also observed for English-Japanese system and Japanese-English system, respectively.

***Keywords:*** statistical machine translation, phrase table, classification model

## 1. Introduction

Statistical machine translation has gained much attention in both academic study and commercial usage for its advancement in the field in recent years. Phrase-based statistical machine translation systems rely on parallel corpora for learning the translation knowledge and translation rules, which are stored in the so-called *phrase table*. The quality of the phrase table is crucial to the translation quality of machine translation systems. Thus, phrase table is the fundamental and vital component in the translation process. A phrase table consists of sequences of words in the source language and sequences of words in the target language, as well as feature scores showing how likely these two sequences are translations of each other. It is usually constructed in two steps: firstly, generating source-to-target and target-to-source word alignments; secondly, extracting bilingual phrase pairs from these alignments through heuristic combination of both directions: this is the grow-diag-final-and heuristic [1].

In Ref. [2], an investigation of the distribution of phrase lengths that are actually used during the decoding process has been conducted. An analysis shows that the majority of phrases used in the translation are phrases of short length. In Ref. [3], an investigation over 10 European language pairs has been conducted to confirm that the majority of phrase pairs used during decoding are phrases with length less than two. The analysis shows that the average percentage over 10 European language pairs is 84%. Therefore, in this paper, we mainly focus on acquisition of bigram pairs from training corpus. Here, bigram pairs consists of 1-to-2, 2-to-1, and 2-to-2 alignments.

As it has been discussed in Refs. [4], [5] that phrase table generated by using the traditional method is not optimal to be used in translation. In this paper, we propose a novel method to detect and extract potential bigram pairs in the training corpus that are not spotted by using the traditional method. To do this, we learn from the decoder. We extract all bigram pairs *actually used* in decoding. We characterize these bigram pairs by computing several features, so that it is possible to classify any new bigram pair as being *potentially useful* or not according to similarity of features. This reduces the problem of extracting potentially useful bigram pairs among all possible bigram pairs to a classification problem.

The focus of this work is to acquire bigram pairs that are not spotted by the traditional phrase extraction method to augment phrase translation table, while at the same time, to improve the translation quality. The merits of the proposed approach are as follows. Firstly, our approach uses the same parallel corpus as the one that is used for training the statistical machine translation system, therefore, no additional corpora are required. Secondly, we use the decoder as the annotator to label candidates in classification model. Thus, it is fully automatic.

The remainder of this paper is organized as follows. Section 2 reviews related works. In Section 3, we present the method to extract potentially useful bigram pairs from the training corpus. Section 4 and Section 5 describe the experimental settings and present the evaluation results. Conclusion and future directions are drawn in Section 6.

## 2. Related Work

In recent years, there are works that have been proposed to deal with phrase tables in statistical machine translation systems. Re-

[1] Graduate School of Information, Production and Systems, Waseda University, Kitakyushu, Fukuoka 808–0135, Japan
[a)] juan.luo@suou.waseda.jp
[b)] yves.lepage@waseda.jp

search on trying to acquire additional data to increase translation coverage have focused on introducing paraphrases, n-grams, and multiword units.

In Ref. [6], paraphrases of unseen source phrases are incorporated into phrase tables by using bitexts. However, their method is particularly pertinent to small corpus and out-of-vocabulary words. Similar to the work presented in Refs. [6], [7] attempted to augment the phrase table with paraphrases. They differ from the previous work in that the paraphrases are derived from monolingual corpus. They also mainly focused on solving the problem of unknown words. In Ref. [8], a method is proposed to augment the phrase tables with paraphrases that are derived from the training parallel corpus. These three works attempted to integrate paraphrases to improve translation coverage and solve the problem of unknown words. Different from the works in Refs. [6], [7], we propose a method to augment phrase tables that does not require additional bilingual or monolingual corpora and our aim is beyond dealing only with unknown words. A method of enlarging n-grams in phrase tables has been reported in Ref. [9], in which "word packing" is used to obtain 1-to-n alignments based on co-occurrence frequencies. Here, our work aim at obtaining not only 1-to-n alignments but also n-to-m phrase pairs. In Ref. [10], collocation segmentation is performed on bilingual corpus to extract n-to-m alignments, which are used to augment phrase tables. However, their experimental results showed no difference in evaluation metric scores. Similar to this work, we also obtain n-to-m phrase pairs from the training corpus to augment the phrase tables. However, we show that our work can achieve improvements in the evaluation scores. Reference [11] proposed a strategy to extract domain bilingual multiword expressions and investigated three methods to integrate these multiword units to phrase tables. It is shown that adding multiword units to an additional phrase table and using multiple phrase tables in Moses can achieve the most improvement among the three methods. Here, in this paper, we also add newly extracted *potentially useful* bigram pairs to a new additional phrase table. In Ref. [12], a hierarchical phrase table combination method is proposed to deal with the data that come from various domains.

The main difference between the above previous works and our work is that we aim at learning from the decoder and seeing how it chooses phrase pairs for translation. We rely on the evidence given by decoder itself.

A number of works have been proposed to solve the problem of alignment and phrase extraction, for example, Refs. [5], [13], [14]. Reference [5] presented an approach to joint phrase alignment and extraction through a hierarchical model using nonparametric Bayesian methods and inversion transduction grammars. In Ref. [13], a general and extensible phrase extraction algorithm is proposed. Reference [14] proposed a sampling-based alignment method to constitute phrase tables. These works are attempting to propose new phrase extraction approaches which are different from the traditional method [1].

Different from these works, we do not invent a new approach that is different from the traditional method. Here we adopt a humble stance on acquiring phrases which is based on the traditional method.

## 3. Classification and Production of New Bigram Pairs

In this section, we present our proposed method to extract potentially useful bigram pairs from training corpus by learning from the decoder. We would like to stress that, by using this method, we are producing new bigram pairs that are not spotted by the traditional method. We also describe features that are used for classification model.

### 3.1 The Method

The procedure consists of two stages: (1) learning the classification model; and (2) producing new bigram pairs. Before presenting the two stages, here, we clarify the terminology that we will use in the text. Since we learn from the decoder about what phrases are *actually used* during translation, we will adopt the term *used* and *not-used* in the first stage. In the second stage, we are producing new phrases, thus we will adopt the term *useful* and *not-useful* to indicate whether the newly generated phrases are potentially useful or not. The detailed process of the method is described as follows (see also **Fig. 1**):

(1) Learning the classification model.

In this stage, we look at the decoder and learn how it chooses the phrase pairs for translation. We formalize this as a classification problem, where a phrase pair is classified as either *used* or *not-used* in the decoding process. To do this, we first extract bigram pairs from a set of sentences in the source language and translation output with the segmentation trace [*1] in the target language. In order to make sure that the target part of these extracted bigram pairs are *correct* translations, they are searched in the corresponding reference sentences (in target language) for confirmation. These extracted and confirmed bigram pairs are labeled as *used* candidates. For such a *used* candidate, we then search in the phrase table and find a *not-used* candidate i.e., a bigram pair in the phrase table that is not used during decoding. There is the case where multiple *not-used* candidates could be found for a *used* candidate. In order to get an equal number of *used* and *not-used* candidates (i.e., get an equal number of positive and negative examples for training classification model), here, we sample and output only one *not-used* candidate to match one *used* candidate. Finally, we define and assign a set of features to the candidates. Given a set of *used* and *not-used* candidates, as well as features, the classification model is constructed.

(2) Producing new bigram pairs.

In this stage, we acquire new bigram pairs (i.e., 1-to-2, 2-to-1, and 2-to-2 pairs). Firstly, a list of bigram pairs are extracted from the training parallel corpus and features are assigned to them. For each of these bigram pairs, the classification model is then employed to predict whether it is *useful* or *not-useful*. Those candidates that are predicted as *useful* and are not found in the baseline phrase translation table are added to augment the phrase translation table. Here, instead of adding the *useful* bigram pairs as new entries to the baseline phrase translation table. we collect these entries to form an additional phrase table. In the experi-

[*1] Option `-t` in Moses. It reveals which phrases were used.

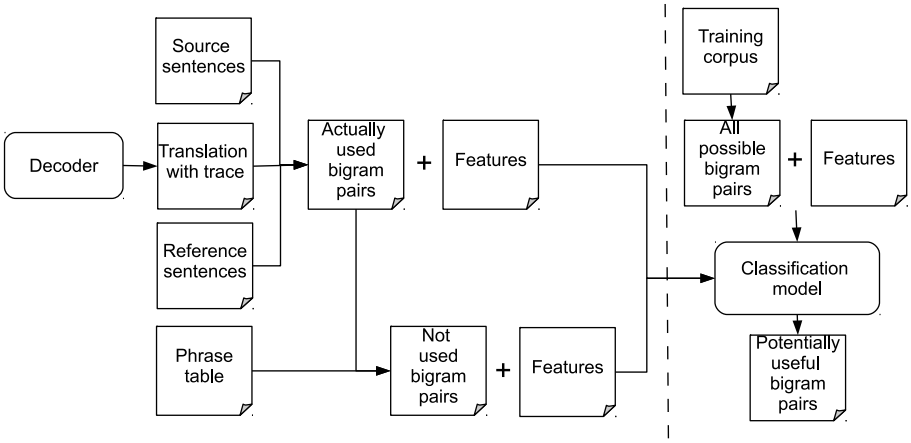



**Fig. 1** Flowchart for the extraction of potentially useful bigram pairs.

ments which are presented in Section 5.2, after obtaining the new phrase table, we run the tuning process again to re-estimate the weights.

### 3.2 An Example

Let us illustrate the method with an example.

(1) Considering the following source sentence, its translation output with segmentation trace and the reference sentence:

source: therefore , the development of instructors and successors was a must .

translation: その␣ため␣|0-0|␣、␣開発␣|1-3|␣の␣|4-4|␣指導␣者␣|5-5|␣と␣後継␣者␣|6-7|␣と␣さ␣れ␣た␣|8-10|␣。

reference: その␣ため␣に␣指導␣者␣・␣後継␣者␣の␣育成␣が␣必須␣で␣あっ␣た␣。

A bigram pair is firstly extracted according to the segmentation trace:

instructors ||| 指導␣者

The target part of this phrase pair is searched in the reference. If it is found, we label it as *used*.

Then, we search for *not-used* candidate in the phrase translation table:

instructors ||| 指導
instructors ||| 指導␣の

Given the same source part of phrase pairs as the *used* candidate, there are two candidates that are not used during decoding. In order to get equal number of positive and negative examples for training classification model, it is sampled and a candidate is extracted and labeled as *not-used*:

instructors ||| 指導␣の

Finally, a pair of candidates is obtained:

*used*: instructors ||| 指導␣者
*not-used*: instructors ||| 指導␣の

(2) Given the training parallel corpus:

source: it was designated a cultural property of the city of kyoto in april 1996 .

target: 1996␣年␣4␣月␣に␣京都␣市␣の␣文化財␣に␣指定␣さ␣れ␣て␣いる␣。

The list of all possible bigram pairs (i.e., 1-to-2, 2-to-1, and 2-to-2 pairs) is extracted:

it ||| 1996␣年
it was ||| 1996
it was ||| 1996␣年
.....

These pairs are assigned with features and predicted by using the classification model. If a newly generated pair is predicted as *useful* and it is not found in the baseline phrase table, such a bigram pair will be added to augment the phrase table:

cultural property ||| 文化財
.....

### 3.3 Features Used for Classification

We used a set of features for classification (see **Table 3**). These features are categorized into global information features and local information features.

For the global information features, we use measures that capture the degree of association of source phrase and target phrase. Translation probabilities and lexical weights [1] (here, computed without alignment) are used. In addition, in word alignment task, the association measures have been proposed to rank and determine if bilingual word pairs are strongly associated with each other. In this work, we use these association measures for global information features: Dice coefficient [15], point-wise mutual information [16], log-likelihood ratio [17], [18], and Pearson's chi-square test [19], [20]. In order to calculate these association measures, two-by-two contingency tables of observed frequencies and expected frequencies are constructed as shown in **Tables 1** and **2**. The cell $n_{11}$ is the number of joint counts of the source and target phrases in the parallel sentences. The cell $n_{12}$ is the number of counts in which the source phrase occurs in the source part

**Table 1** Contingency table for observed frequencies.

| | Target phrase | ¬Target phrase | Total |
|---|---|---|---|
| Source phrase | $n_{11}$ | $n_{12}$ | $S_1$ |
| ¬Source phrase | $n_{21}$ | $n_{22}$ | $S_2$ |
| Total | $T_1$ | $T_2$ | N |

**Table 2** Contingency table for expected frequencies.

| | Target phrase | ¬Target phrase | Total |
|---|---|---|---|
| Source phrase | $m_{11} = \frac{S_1 \times T_1}{N}$ | $m_{12} = \frac{S_1 \times T_2}{N}$ | $S_1$ |
| ¬Source phrase | $m_{21} = \frac{T_1 \times S_2}{N}$ | $m_{22} = \frac{T_2 \times S_2}{N}$ | $S_2$ |
| Total | $T_1$ | $T_2$ | N |

…… 文化 財 ……  …… cultural assets ……

…… 有形 文化 財 ……  …… tangible cultural property ……

…… 無形 文化 財 ……  …… intangible cultural property ……

$$\phi(\text{cultural property} \mid \text{文化 財}) = \frac{2}{3} = 0.67$$

**Fig. 2** Example of computation of translation probability of a phrase pair.

and the target phrase does not occur in the target part. The cell $n_{21}$ is the number of counts in which the target phrase occurs in the target part and the source phrase does not occur in the source part. The cell $n_{22}$ is the number of counts in which neither the source phrase or the target phrase occur in the parallel sentences. $N$ is the number of parallel sentences.

For the local information features, we use measures that capture the unithood and degree of association of each word in a monolingual phrase. Language model [21] and generalized dice coefficient [22] are used.

- Global information features

  Translation probabilities:

  Given the phrase pairs, the translation probability is estimated by the relative frequency:

$$\phi(\bar{f}|\bar{e}) = \frac{count(\bar{f}, \bar{e})}{count(\bar{e})} \tag{1}$$

  where $count(\bar{f}, \bar{e})$ is considered here as the number of times that the source phrase $\bar{f}$ and the target phrase $\bar{e}$ are found to co-occur in the same line of the corpus (see **Fig. 2**).

  Lexical weights:

  Given a phrase pair $\bar{f}, \bar{e}$, and a word alignment $a$ between the source word positions $i = 1, \ldots, n$ and the target word positions $j = 0, 1, \ldots, m$, the lexical weight in Ref. [1] is computed as:

$$p_w(\bar{f}|\bar{e}, a) = \prod_{i=1}^{n} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall (i,j) \in a} w(f_i|e_j) \tag{2}$$

  Here, we compute the score as the following equation and it is computed without alignment (see **Fig. 3**):

$$p_w(\bar{f}|\bar{e}) = \prod_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} w(f_i|e_j) \tag{3}$$

  Dice Coefficient [15]:

  The Dice coefficient measures how often two phrase pairs co-occur in their respective sentences.

$$Dice(s,t) = \frac{2 \times n_{11}}{S_1 + T_1} \tag{4}$$

| | cultural | property |
|---|---|---|
| NULL | | |
| 文化 | ■ | |
| 財 | | ■ |

According to equation (2) [1]

| | cultural | property |
|---|---|---|
| NULL | ■ | ■ |
| 文化 | ■ | ■ |
| 財 | ■ | ■ |

According to equation (3)

**Fig. 3** Computation of lexical weights according to Eq. (2) (left) and Eq. (3) (right).

  Point-wise Mutual Information [16]:

  This measure counts the co-occurrence frequency of phrase pairs. It is considered as perhaps the most widely used measure in extraction of collocations [23].

$$PMI(s,t) = \log_2 \frac{n_{11}}{m_{11}} \tag{5}$$

  Log-likelihood Ratio [17], [18]:

  It is noted in Ref. [16] that the advantage of this measure is that it takes into consideration of all the cases where the phase pairs co-occur or do not co-occur in their respective source and target lines of the corpus. It is also pointed out in Ref. [23] that Log-likelihood Ratio is appropriate for the case where the data are sparse.

$$LLR(s,t) = \sum_{i,j} n_{ij} \log \frac{n_{ij} N}{S_i T_j} \tag{6}$$

  Pearson's Chi-square test [19], [20]:

  This measure are usually used to test if two hypotheses co-occur coincidentally. The higher the score, the more they are dependent on each other [24].

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{7}$$

- Local information features

  Generalized Dice Coefficient [22]:

  This measure computes the association of an arbitrary n-gram and measures its cohesion.

$$GDC(T) = \frac{|T| \times \log_{10} f(T) \times f(T)}{\sum_{w_i \in T} f(w_i)} \tag{8}$$

  where $|T|$ is the length of the phrases in words in phrase $T$, $f(T)$ is the frequency of phrase $T$, and $f(w_i)$ is the frequency of word $w_i$.

  Language model [21]:

  Here, the SRILM toolkit [21] is used to build language models from the training data and the well-known Kneser-Ney smoothing is used [*2].

## 4. Experimental Setting

In this section, we present the experiments on two language pairs in both directions: Chinese-English and Japanese-English. We would like to stress that the bigram pairs comprise 1-to-2, 2-to-1, and 2-to-2 alignments.

[*2] Here, we query the language model by using `mosesdecoder/bin/query` in Moses.

**Table 3** Summary and characterization of features.

| | Monolingual | Bilingual | |
|---|---|---|---|
| | | One-directional | Two-directional |
| Language model | 2 | | |
| Generalized Dice Coefficient | 2 | | |
| Translation probabilities | | 2 | |
| Lexical weights | | 2 | |
| Dice Coefficient | | | 1 |
| Point-wise Mutual Information | | | 1 |
| Log-likelihood Ratio | | | 1 |
| Pearson's Chi-square test | | | 1 |

**Table 4** Statistics on the datasets (M = million).

| | | Japanese | English | Chinese | English |
|---|---|---|---|---|---|
| train | sentences | 300,000 | | 300,000 | |
| | tokens | 5.13M | 5.40M | 7.78M | 8.38M |
| | types | 90,593 | 129,934 | 66,078 | 55,674 |
| dev. | sentences | 500 | | 500 | |
| | tokens | 14,433 | 13,986 | 16,653 | 18,749 |
| | types | 3,038 | 2,790 | 1,731 | 1,966 |
| test | sentences | 1,000 | | 1,000 | |
| | tokens | 22,339 | 22,781 | 31,122 | 34,729 |
| | types | 4,212 | 3,918 | 2,353 | 2,786 |
| for | sentences | 1,000 | | 1,000 | |
| classification | tokens | 15,932 | 15,897 | 32,469 | 35,690 |
| | types | 2,653 | 2,784 | 2,467 | 2,819 |

### 4.1 Experimental Setup

Standard statistical machine translation systems were built by using the conventional pipeline: the Moses toolkit [25], MERT (Minimum Error Rate Training) [26] to tune the parameters, the SRI Language Modeling (SRILM) toolkit [21] to build a 5-gram target language model with Kneser-Ney smoothing, and GIZA++ [27] to generate word alignment. The maximum length of phrase pairs in phrase tables is set to 7 (the default phrase length in Moses). The distortion limit is 6. For the evaluation of translations, four standard automatic evaluation metrics were used: BLEU [28], NIST [29], WER [30], and TER [31].

To construct a classification model, we use Support Vector Machine (SVM). In this work, we used LIBSVM [32]. The radial basis function kernel and 5-fold cross-validation are used.

### 4.2 Experimental Dataset

For the Chinese-English task, the train, development, and test sets were extracted from the MultiUN corpus [33]. The Stanford Chinese Word Segmenter [34] is used to segment the Chinese sentences. For the Japanese-English task, we used a sample extracted from the KFTT data [35]. For both Chinese-English and Japanese-English tasks, we used a training set of 300,000 sentences. The development sets contain 500 sentences, and 1,000 sentences are used for test sets. A detailed description of the datasets is given in **Table 4**. We also used an additional 1,000 sentences from the above corpora of the same domain to build a classification model. We would like to stress that these sentences are different from the datasets (i.e., train, dev., and test data) that are used in machine translation tasks.

## 5. Experimental Results

In this section, we present the evaluation results of two experiments. In the first experiment, we evaluate the performance of classifier on predicting the bigram pairs. In the second experiment, the performance of spotting potential bigram pairs to augment phrase translation tables in machine translation systems is evaluated.

**Table 5** The data for classification test. P is the number of positive candidates. N is the number of negative candidates.

| | Train | Test |
|---|---|---|
| Japanese-English | 1,400 (P: 700; N: 700) | 200 (P: 100; N: 100) |
| English-Japanese | 1,400 (P: 700; N: 700) | 200 (P: 100; N: 100) |
| Chinese-English | 980 (P: 490; N: 490) | 200 (P: 100; N: 100) |
| English-Chinese | 1,100 (P: 550; N: 550) | 200 (P: 100; N: 100) |

**Table 6** Accuracy of the classifier on the test sets for individual features and feature combination.

| | Ja-En | En-Ja | Zh-En | En-Zh |
|---|---|---|---|---|
| Prob. | 86.5% | 78.0% | 72.5% | 68.5% |
| LW | 81.5% | 79.5% | 68.0% | 70.5% |
| Dice | 63.5% | 69.0% | 65.5% | 72.0% |
| PMI | 47.5% | 59.5% | 62.5% | 67.5% |
| LLR | 75.0% | 75.0% | 71.0% | 79.5% |
| $\chi^2$ | 67.0% | 67.0% | 68.0% | 78.0% |
| GDC | 74.0% | 72.5% | 76.0% | 69.5% |
| LM | 74.0% | 71.5% | 61.5% | 62.0% |
| All | **91.0%** | **88.0%** | **88.5%** | **88.5%** |

### 5.1 Classification Test

In this experiment, we extracted bigram pairs from a set of 1,000 parallel sentences (see Table 4) (i.e., source sentences and their translation output with segmentation trace). These bigram pairs are sampled and splitted into train and test sets for experiment (see **Table 5**). It should be noted that the number of candidates for training are different for the systems of different language pairs. This is because that, by looking at the segmentation trace, the length of translation unit and choice of translation phrases are different for each language pair. All the datasets contain an equal number of positive and negative candidates. We used the built-in accuracy measure of LIBSVM to evaluate the performance of classifier on the test set:

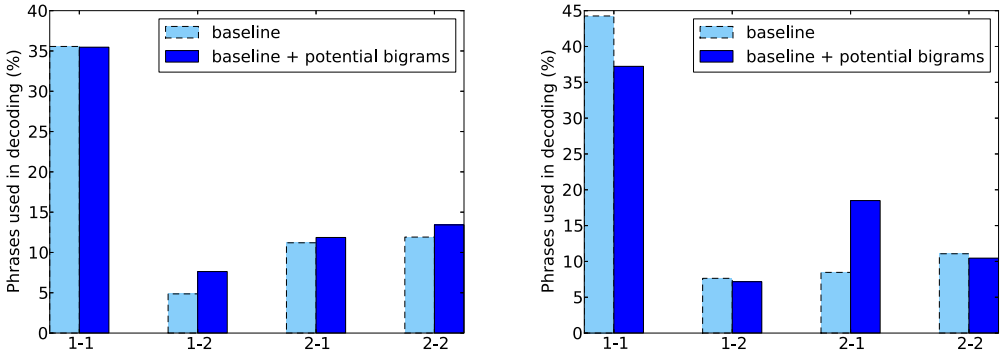
$$accuracy = \frac{C_{tp} + C_{tn}}{C} \tag{9}$$

where $C_{tp}$ is the counts of true-positive and $C_{tn}$ is the counts of true-negative. $C$ is the total counts of candidates.

The evaluation results are shown in **Table 6**. From the table we can see that the translation probabilities (Prob.) and lexical weights (LW) features are the most efficient among all features for Japanese-English and English-Japanese tasks. However, this is not so much the case for the language pair Chinese-English. The most informative features that are observed for Chinese-English and English-Chinese are generalized dice coefficient (GDC) and log-likelihood ratio (LLR). By using a combination of all features, the best performance is shown for Japanese-English, where an accuracy of 91% is observed. For the other tasks, the accuracy ranges from 88.0% to 88.5%. The results have shown that using all the features allowed us to achieve the best performance.

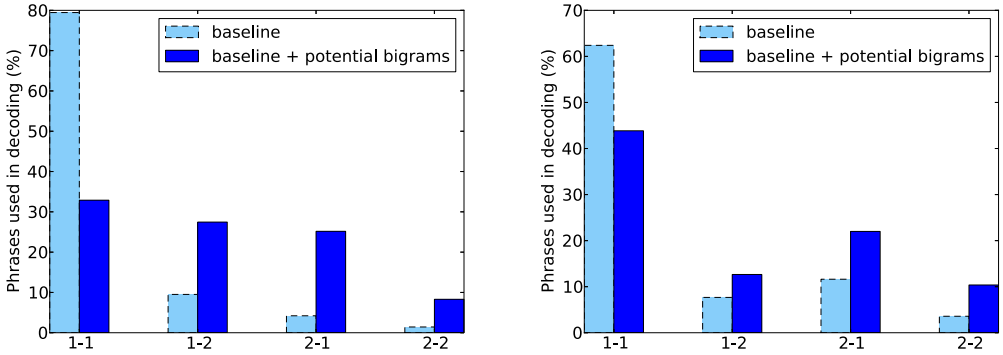
### 5.2 Machine Translation Test

In this experiment, we evaluate the proposed method in statistical machine translation tasks. The evaluation results are shown in




(a) ja-en (b) en-ja

**Fig. 4** Distribution of phrases used during decoding, length = 2.




(a) zh-en (b) en-zh

**Fig. 5** Distribution of phrases used during decoding, length = 2.

**Table 7** Evaluation results (BL: baseline; 2-g: potentially useful bigram pairs) (training data: 300,000 lines).

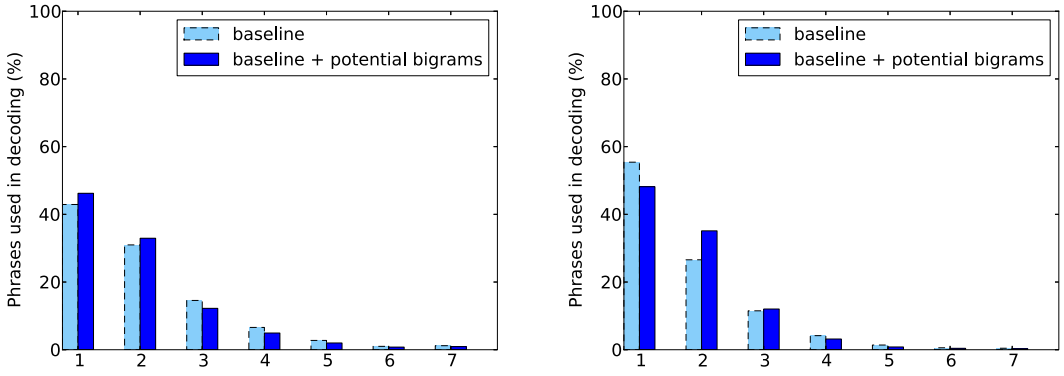
| | | BLEU | NIST | WER | TER |
|---|---|---|---|---|---|
| Japanese-English | BL | 18.54 | 5.7621 | 0.6624 | 0.6943 |
| | BL + 2-g | **18.94** | **5.7711** | 0.6757 | 0.7005 |
| English-Japanese | BL | 20.50 | 5.5471 | 0.7402 | 0.7450 |
| | BL + 2-g | **21.44** | **5.6935** | **0.7238** | **0.7196** |
| Chinese-English | BL | 21.36 | 5.0683 | 0.6355 | 0.6855 |
| | BL + 2-g | **28.97** | **6.1215** | **0.5710** | **0.6163** |
| English-Chinese | BL | 18.51 | 5.0897 | 0.6585 | 0.7306 |
| | BL + 2-g | **26.16** | **6.2701** | **0.6165** | **0.6461** |

**Table 7**. Significant improvements are achieved for the Chinese-English and English-Chinese systems, in which an increase of 7.61 and 7.65 BLEU point are observed respectively. For the English-Japanese task, by comparing with the baseline, the proposed approach outperforms in all four evaluation metrics with an increase of 0.94 in BLEU points. For the Japanese-English task, extending potential bigram pairs to phrase translation table leads to improvement in BLEU and NIST scores, however, decreases are observed for the metrics WER and TER.

We analyzed the distribution of phrases used during decoding for the baseline systems and the systems with *potentially useful* bigram pairs (see **Figs. 4** and **5**). In order to examine the effect of the proposed method, here we only show the distribution of phrases in which their lengths are less than two. From the figures, it can be seen that, by comparing with the baseline, the more *potentially useful* bigram pairs (i.e., 1-to-2, 2-to-1, and 2-to-2 pairs) are used during decoding, the larger the improvements in evaluation scores are observed for the systems with *potentially useful* bigram pairs. The greatest increase in the number of bigram pairs is observed for the language pair Chinese-English, where the percentage increases from 9.51%, 4.21%, 1.43% to 27.46%, 25.17%, 8.30% for 1-to-2, 2-to-1, and 2-to-2, respectively. In total, an increase of 46.56% is observed for these pairs. This shows that the addition of *potentially useful* bigram pairs is beneficial to this machine translation system.

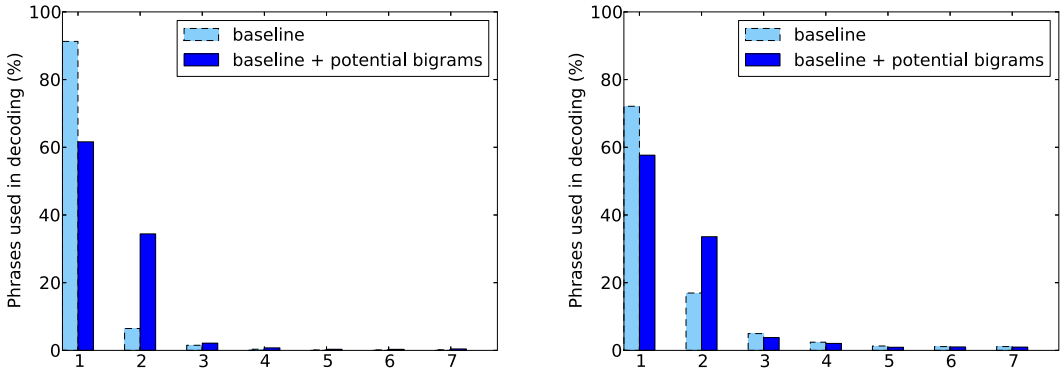
In **Figs. 6** and **7**, the full distribution of phrases with length less than seven are shown. From the figures, we can see that the distributions of phrases between two language pairs, that is, Japanese-English and Chinese-English, are different. In the baseline systems, the percentage of phrases of length one and two are 73.88% and 81.98% for Japanese-English and English-Japanese, respectively. The number of phrases of length three represents 14.56% and 11.50%. In the baseline Chinese-English and English-Chinese systems, the percentage of phrases of length one and two are 97.74% and 89.07%. The percentage of phrases of length three are 1.51% and 4.96%, respectively. Here, we can see that there is a big difference in the number of phrases of length three between two language pairs. There are more phrases of length three are employed in translation of Japanese than Chinese sentences (see also **Figs. 8** and **9**). This may explain the difference of the impact of the method in BLEU scores between two language pairs, i.e., Chinese-English and Japanese-English, since we mainly focused on acquisition of bigram pairs in this work.

We also analyzed how many the newly generated potentially useful bigram pairs are used in translation. This is shown in **Ta-**




(a) ja-en (b) en-ja

**Fig. 6** Distribution of source phrases used during decoding by length 7. Japanese-English is on the left and the source is Japanese. English-Japanese is on the right and the source is English.




(a) zh-en (b) en-zh

**Fig. 7** Distribution of source phrases used during decoding by length 7. Chinese-English is on the left and the source is Chinese. English-Chinese is on the right and the source is English.

**Table 8** Analysis of number of phrase pairs used in translations that come from the newly generated entries (BL + 2-g).

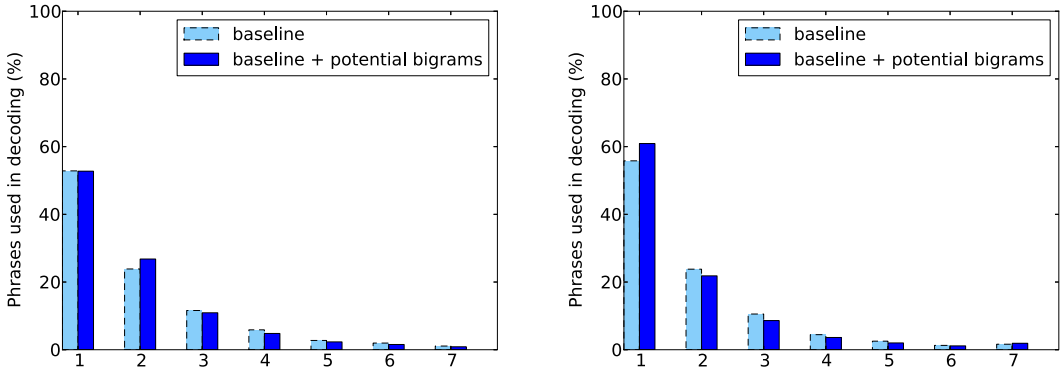
| | | Phrases | From New | Percentage |
|---|---|---|---|---|
| Japanese-English | total | 11,790 | 317 | 2.69% |
| | 1-to-2 | 900 | 206 | 22.89% |
| | 2-to-1 | 1,398 | 76 | 5.44% |
| | 2-to-2 | 1,582 | 35 | 2.21% |
| English-Japanese | total | 12,960 | 390 | 3.01% |
| | 1-to-2 | 929 | 104 | 11.19% |
| | 2-to-1 | 2,396 | 222 | 9.27% |
| | 2-to-2 | 1,355 | 64 | 4.72% |
| Chinese-English | total | 21,260 | 10,605 | 49.88% |
| | 1-to-2 | 5,838 | 4,671 | 80.01% |
| | 2-to-1 | 5,351 | 4,397 | 82.17% |
| | 2-to-2 | 1,765 | 1,537 | 87.08% |
| English-Chinese | total | 21,455 | 6,628 | 30.89% |
| | 1-to-2 | 2,714 | 1,796 | 66.18% |
| | 2-to-1 | 4,721 | 3,193 | 67.63% |
| | 2-to-2 | 2,225 | 1,639 | 73.66% |

**Table 9** Examples of newly generated bigram pairs which were used in translations.

| | New bigram pairs |
|---|---|
| Japanese-English | 大 合併 ||| great merger |
| | 製鉄 所 ||| iron factory |
| | 移築 ||| was relocated |
| English-Japanese | single-track section ||| 単線 区間 |
| | regular train ||| 定期 列車 |
| | military advisors ||| 軍事 顧問 |
| Chinese-English | 特别 报告员 ||| special rapporteur |
| | 不 相容 ||| incompatible |
| | 人道主义法 ||| humanitarian law |
| English-Chinese | technical advice ||| 技术 咨询 |
| | inhuman ||| 不 人道 |
| | but also ||| 而且 还 |

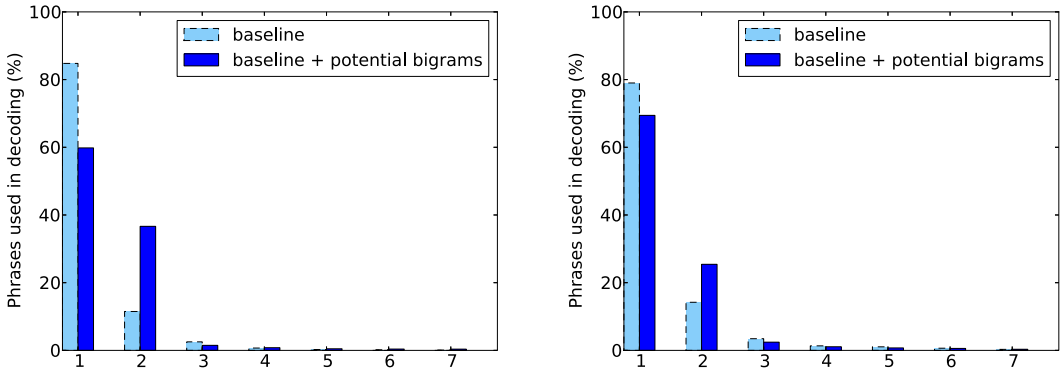
**ble 8** (see **Table 9** for examples). From the table, we can see that only a small number of newly generated bigram pairs are used in translations in Japanese-English and English-Japanese systems. The percentage is 2.69% and 3.01%. This may be the reason for a slight increase in BLEU scores for these machine translation systems. As for Chinese-English and English-Chinese systems, there are 49.88% and 30.89% of the bigram pairs that are used in translation are coming from the newly generated entries. This shows that the proposed method is indeed producing new bigram pairs which contribute to the significant improvement in translation results for Chinese-English and English-Chinese systems.

In addition, we experimented on increasing the data size. For the Chinese-English tasks, we increased the number of sentences for training from 300,000 to 400,000. The number of sentences for development, test and classification are the same. The evaluation results are shown in **Table 10**. From the table we can see that a significant improvement is also achieved. An improvement of 6.18 and 5.4 BLEU points are observed for Chinese-English and English-Chinese tasks.




(a) ja-en (b) en-ja

**Fig. 8** Distribution of target phrases used during decoding by length 7. Japanese-English is on the left and the target is English. English-Japanese is on the right and the target is Japanese.




(a) zh-en (b) en-zh

**Fig. 9** Distribution of target phrases used during decoding by length 7. Chinese-English is on the left and the target is English. English-Chinese is on the right and the target is Chinese.

**Table 10** Evaluation results (BL: baseline; 2-g: potentially useful bigram pairs) (training data: 400,000 lines).

| | | BLEU | NIST | WER | TER |
|---|---|---|---|---|---|
| Chinese-English | BL | 23.81 | 5.3614 | 0.6051 | 0.6693 |
| | BL + 2-g | **29.99** | **6.2938** | **0.5547** | **0.6064** |
| English-Chinese | BL | 20.36 | 5.2889 | 0.6345 | 0.7143 |
| | BL + 2-g | **25.76** | **6.2594** | **0.6070** | **0.6436** |

## 6. Conclusion

In this paper, we proposed a novel method to extract potentially useful phrase pairs that are not output by the traditional phrase extraction heuristic to augment the phrase table. We extracted potentially useful bigram pairs from the training corpus, which is approached as a classification problem. A set of features is defined to capture the bilingual association of the source and target phrases, as well as the monolingual association between words in each language. Experiments were conducted to assess the performance of the proposed method. A statistically significant increase of 7.65 and 7.61 BLEU points were achieved in the English-Chinese and Chinese-English tasks, respectively. A slight improvement of 0.94 and 0.4 BLEU points for the Japanese-English and English-Japanese was also observed.

We believe that this approach can be extended and further improved in a number of ways. In this paper, we only considered bigram pairs. The extension to longer n-grams (e.g., trigrams, tetragrams) should be inquired. The inclusion of more features for classification should also be inquired, such as C-value [36]. Since all possible bigram pairs from the source and the target sentences are considered for classification, the result of extraction is still noisy. The reduction of noise should be investigated. We will also experiment on setting a threshold according to the distance from the hyper plane.

### References

[1] Koehn, P., Och, F.J. and Marcu, D.: Statistical Phrase-Based Translation, *Proc. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, pp.48–54 (2003).

[2] Ayan, N.F. and Dorr, B.J.: Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT, *Proc. 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp.9–16 (2006).

[3] Luo, J. and Lepage, Y.: A comparison of association and estimation approaches to alignment in word-to-word translation, *Proc. 10th International Symposium on Natural Language Processing*, Phuket, Thailand, pp.181–186 (2013).

[4] DeNero, J. and Klein, D.: Discriminative Modeling of Extraction Sets for Machine Translation, *Proc. 48th Annual Meeting of the Association of Computational Linguistics*, Uppsala, Sweden, pp.1453–1463 (2010).

[5] Neubig, G., Watanabe, T., Sumita, E., Mori, S. and Kawahara, T.: An Unsupervised Model for Joint Phrase Alignment and Extraction, *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp.632–641 (2011).

[6] Callison-Burch, C., Koehn, P. and Osborne, M.: Improved statistical

machine translation using paraphrases, *Proc. Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, New York, pp.17–24 (2006).
[7] Marton, Y., Callison-Burch, C. and Resnik, P.: Improved statistical machine translation using monolingually-derived paraphrases, *Proc. Conference on Empirical Methods on Natural Language Processing*, Singapore, pp.381–390 (2009).
[8] Fujita, A. and Carpuat, M.: FUN-NRC: Paraphrase-augmented Phrase-based SMT Systems for NTCIR-10 PatentMT, *Proc. 10th NTCIR*, pp.327–334 (2013).
[9] Ma, Y., Stroppa, N. and Way, A.: Bootstrapping Word Alignment via Word Packing, *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp.304–311 (2007).
[10] Henríquez, Q.A.C., Costa-jussà, R.M., Daudaravicius, V., Banchs, E.R. and Mariño, B.J.: Using collocation segmentation to augment the phrase table, *Proc. Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden, pp.98–102 (2010).
[11] Ren, Z., Lü, Y., Cao, J., Liu, Q. and Huang, Y.: Improving statistical machine translation using domain bilingual multiword expressions, *Proc. Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Suntec, Singapore, pp.47–54 (2009).
[12] Zhu, C., Watanabe, T., Sumita, E. and Zhao, T.: Hierarchical Phrase Table Combination for Machine Translation, *Proc. 51st Annual Meeting of the Association of Computational Linguistics*, Sofia, Bulgaria, pp.802–810 (2013).
[13] Ling, W., Luis, T., Graca, J., Coheur, L. and Trancoso, I.: Towards a General and Extensible Phrase-Extraction Algorithm, *Proc. International Workshop on Spoken Language Translation*, France, pp.313–320 (2010).
[14] Lardilleux, A. and Lepage, Y.: Sampling-based multilingual alignment, *Proc. International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp.214–218 (2009).
[15] Dice, L.R.: Measures of the amount of ecologic association between species, *Ecology*, Vol.26, No.3, pp.297–302 (1945).
[16] Kobdani, H., Fraser, A. and Schütze, H.: Word Alignment by Thresholded Two-Dimensional Normalization, *Proc. MT Summit XII*, Ottawa, Ontario, Canada, pp.260–267 (2009).
[17] Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol.19, No.1, pp.61–74 (1993).
[18] Moore, R.C.: On Log-Likelihood-Ratios and the Significance of Rare Events, *Proc. Conference on Empirical Methods on Natural Language Processing*, Barcelona, Spain, pp.333–340 (2004).
[19] Gale, W.A. and Church, K.W.: Identifying Word Correspondence in Parallel Texts, *Proc. Workshop on Speech and Natural Language*, Pacific Grove, California, pp.152–157 (1991).
[20] Hoang, C., Le, C.-A. and Pham, S.-B.: Improving the Quality of Word Alignment by Integrating Pearson's Chi-Square Test Information, *Proc. International Conference on Asian Language Processing*, Hanoi, pp.121–124 (2012).
[21] Stolcke, A.: SRILM—An extensible language modeling toolkit, *Proc. 7th International Conference on Spoken Language Processing*, Denver, Colorado, pp.901–904 (2002).
[22] Park, Y., Byrd, R.J. and Boguraev, B.K.: Automatic Glossary Extraction: Beyond Terminology Identification, *Proc. International Conference on Computational Linguistics*, Taipei, Taiwan, pp.1–7 (2002).
[23] Pereira, L., Strafella, E. and Matsumoto, Y.: Collocation or Free Combination? Applying Machine Translation Techniques to Identify Collocations in Japanese, *Proc. 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, pp.736–739 (2014).
[24] Chang, B. and Han, D.: Enhancing domain portability of Chinese segmentation model using chi-square statistics and bootstrapping, *Proc. Empirical Methods in Natural Language Processing*, Massachusetts, USA, pp.789–798 (2010).
[25] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E.: Moses: Open source toolkit for statistical machine translation, *Proc. 45th Annual Meeting on Association for Computational Linguistics*, Prague, Czech Republic, pp.177–180 (2007).
[26] Och, F.J.: Minimum error rate training in statistical machine translation, *Proc. 41st Annual Meeting on Association for Computational Linguistics*, Sapporo, Japan, pp.160–167 (2003).
[27] Och, F.J. and Ney, H.: A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, Vol.29, pp.19–51 (2003).
[28] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, *Proc. 40th Annual Meeting of the Association of Computational Linguistics*, Philadelphia, pp.311–318 (2002).
[29] Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, *Proc. 2nd International Conference on Human Language Technology Research*, San Diego, pp.138–145 (2002).
[30] Nießen, S., Och, F.J., Leusch, G. and Ney, H.: An evaluation tool for machine translation: Fast evaluation for machine translation research, *Proc. 2nd International Conference on Language Resources and Evaluation*, Athens, pp.39–45 (2000).
[31] Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J.: A study of translation edit rate with targeted human annotation, *Proc. 7th Association for Machine Translation in the Americas*, Cambridge, Massachusetts, pp.223–231 (2006).
[32] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Trans. Intelligent Systems and Technology*, Vol.2, No.27, pp.1–27 (2011).
[33] Eisele, A. and Chen, Y.: MultiUN: A Multilingual Corpus from United Nation Documents, *Proc. 7th Language Resources and Evaluation Conference*, La Valletta, Malta, pp.2868–2872 (2010).
[34] Tseng, H., Chang, P., Andrew, G., Jurafsky, D. and Manning, C.: A Conditional Random Field Word Segmenter for Sighan Bakeoff, *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, Jeju Island, Korea, pp.168–171 (2005).
[35] Neubig, G.: The Kyoto Free Translation Task, available from ⟨http://www.phontron.com/kftt⟩ (2011).
[36] Frantzi, K.T. and Ananiadou, S.: Extracting Nested Collocations, *Proc. 16th International Conference on Computational linguistics*, Copenhagen, Denmark, pp.41–46 (1996).

**Juan Luo** is currently a Ph.D. candidate in the Example-Based Machine Translation/NLP laboratory at Waseda University. Her research interests include machine translation and natural language processing.

**Yves Lepage** received his D.E.A. and Ph.D. degrees in 1989 from Grenoble university, France, in GETA under the supervision of Professor Vauquois and Professor Boitet. After a post-doctorate at ELSAP, university of Caen and EDF, Paris, he joined ATR labs, Japan, where he worked as an invited researcher and a senior researcher until 2006. In 2003 he got the habilitation for his habilitation thesis entitled "Of the kind of analogies that renders an account of commutations in linguistics." In October 2006, he got the qualification for full professorship from the National Board of French Universities in both linguistics and computer science and became full professor at the University of Caen Basse-Normandie in October 2006. He joined Waseda University, graduate school of Information, Production and Systems in April 2010. His research interests are in Natural Language Processing, Machine Translation, and in particular Example-Based Machine Translation. He is a member of the French and the Japanese Natural Language Processing Associations. He is a member of the board of the French Natural Language Processing Association, ATALA, and one of the four editors-in-chief of the French journal on Natural Language Processing, TAL.