Leveraging Additional Resources for Improving Statistical Machine Translation on Asian Low-Resource Languages

HAI-LONG TRIEU, DUC-VU TRAN, Japan Advanced Institute of Science and Technology
ASHWIN ITTOO, University of Liège
LE-MINH NGUYEN, Japan Advanced Institute of Science and Technology

Phrase-based machine translation systems require large bilingual corpora for training. Nevertheless, such large bilingual corpora are unavailable for most language pairs in the world, which cause a bottleneck for the development of machine translation. For the Asian language pairs: Japanese, Indonesian, Malay paired with Vietnamese, they are also not excluded from the case, in which there are no large bilingual corpora on these low-resource language pairs. Furthermore, although the languages are widely used in the world, there is no prior work of machine translation, which causes an issue for the development of machine translation on these languages. In this paper, we conducted an empirical study of leveraging additional resources to improve machine translation for the Asian low-resource language pairs: translation from Japanese, Indonesian, and Malay to Vietnamese. We propose an innovative approach lies in two strategies of building bilingual corpora from comparable data and phrase pivot translation on existing bilingual corpora of the languages paired with English. Bilingual corpora were built from Wikipedia bilingual titles to enhance bilingual data for the low-resource languages. Additionally, we introduced a combined model of the additional resources to create an effective solution to improve machine translation on the Asian low-resource languages. Experimental results show the effectiveness of our systems with the improvement of +2 to +7 BLEU points. This work contributes to the development of machine translation on low-resource languages, especially opening a promising direction for the progress of machine translation on the Asian language pairs.

CCS Concepts: •**Computing methodologies** → **Machine translation; Language resources;** *Lexical semantics;*

Additional Key Words and Phrases: Statistical machine translation, pivot methods, sentence alignment, semantic similarity, low-resource languages

ACM Reference Format:
Hai-Long Trieu, Duc-Vu Tran, Ashwin Ittoo, Le-Minh Nguyen, 2017. Leveraging Additional Resources for Improving Machine Translation on Asian Low-Resource Languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 9, 4, Article 39 (March 2017), 23 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Phrase-based statistical machine translation (SMT) requires large bilingual corpora for training machine translation models. Given large bilingual corpora, SMT systems achieve high performance in many language pairs [Bojar et al. 2013]. Nevertheless, such large bilingual corpora are unavailable for most language pairs except for a few languages such as several European languages, Arabic, and Chinese [Wang et al.

Author's addresses: Hai-Long Trieu, Duc-Vu Tran, Le-Minh Nguyen, School of Information Science, Japan Advanced Institute of Science and Technology, Asahidai 1-1, Nomi, Ishikawa, Japan; Emails: {trieulh, vu.tran, nguyenml}@jaist.ac.jp. Ashwin Ittoo, QUANTOM (Centre for Quantitative Methods and Operations Management), HEC Liège, University of Liège, Rue Louvrex 14, Liège 4000, Belgium; Email: Ashwin.Ittoo@ulg.ac.be

DOI: 0000001.0000001

2016]. For such language pairs called low-resource languages, in which there are fewer than million bilingual sentences, the performance of SMT systems degrade greatly because of the less reliable translation rules, poorer rule coverage, and of the sparse word and phrase counts that define parameters of SMT models. Therefore, improving performance for low-resource languages becomes an important task for machine translation (MT).

Several solutions have been proposed in previous work to improve MT on low-resource languages by leveraging additional resources. The first strategy is to extract parallel sentences from comparable data. Resnik [Resnik 1999] introduces a method to automatically find parallel documents on the web, which is applied for extracting a French-English corpus. Utiyama and Isahara [Utiyama and Isahara 2003] aligned a large Japanese-English corpus using some manual involvement for matching documents. In a recent work, Chu et al., [Chu et al. 2014] present a work of constructing a Chinese-Japanese parallel corpus from Wikipedia based on a filtering method of sentence extraction using feature sets for classification. The second strategy is to utilize monolingual corpora for learning a phrase-based translation model as in the work of [Ravi and Knight 2011; Nuhn et al. 2012; Saluja et al. 2014]. Another strategy is to utilize existing bilingual corpora using pivot methods, which use pivot languages as bridge for translation models trained on the source-pivot and pivot-target bilingual corpora [De Gispert and Marino 2006; Cohn and Lapata 2007; Utiyama and Isahara 2007; Wu and Wang 2007]. The strategies have been shown to be effectiveness on several low-resource languages. However, Asian low-resource language pairs, such as Japanese, Indonesian, Malay, and Vietnamese have been largely ignored.

In this work, our goal is to improve MT on several Asian low-resource languages: Japanese, Indonesian, Malay, and Vietnamese. The motivation is that although the languages are ranked as some of the most widely used languages in the world: Japanese (rank 9), Indonesian (rank 12), and Vietnamese (rank 17) [Weber 2008], there is no prior work of MT on these language pairs. To date, no translation system has been proposed for these languages. Additionally, there are no large bilingual corpora on the language pairs, which cause a bottleneck for MT.

In this paper, we propose an innovative approach for improving MT on the Asian low-resource languages. The main innovation in our approach lies in two strategies: building bilingual corpora from comparable data and exploiting existing bilingual corpora using the pivot methods. For the first strategy, we built parallel corpora for the Asian language pairs from Wikipedia articles. Parallel titles of Wikipedia articles were used to collect bilingual Wikipedia articles. Then, a powerful sentence alignment method of length-based and word-based [Moore 2002] was used to extract parallel sentences. For the second strategy, we exploited existing bilingual corpora to improve MT on the Asian language pairs using the well-known phrase pivot translation approach [Wu and Wang 2007; Utiyama and Isahara 2007]. In order to create an effective solution for improving MT on the Asian low-resource languages, we introduce a combined model to take advantage of the additional resources most effectively. Experimental results show that the strategies have been applied successfully and show the effectiveness of our extracted Wikipedia corpus as well as the phrase pivot translation on the Asian language pairs. For the combined model, our systems achieved the significant improvement from +2 to +7 BLEU points by leveraging the additional resources.

The contribution of this paper is in the following aspects:

(1) Propose and develop an MT system for the Asian low-resource languages: translation from Japanese, Malay, and Indonesian to Vietnamese, in which there is no prior work of MT on such language pairs, especially in the task of improving MT

on low-resource languages although the languages are in the most widely used in the world
(2) Build new bilingual corpora for the Asian low-resource languages, which showed the significant improvement for MT on the low-resource languages
(3) Examine and investigate the well-known strategies for improving MT on the Asian low-resource languages: building bilingual corpora and exploiting existing corpora
(4) Introduce a combined model which showed significant improvement and create an effective solution for MT on the Asian low-resource languages
(5) Provide data sets as well as solutions, which can be useful for further development of improving MT on the Asian low-resource languages; all of the source code, data sets as well as our extracted corpus are available at the repository.[1]

The structure of this paper is organized as follows. In Section 2, we review previous work that is related to our research including of building bilingual corpora from comparable data, phrase pivot translation, and improving MT on low-resource languages. In Section 3, methods used in our research are described, which are the procedure and approach of building bilingual corpora from Wikipedia, the phrase pivot translation, and our combined model based on a linear interpolation with perplexity minimization. Experiments are presented in Section 4, in which we describe in details the configuration of our systems, data sets, and experimental results. Further analyses were discussed in Section 5, which provides an in-depth investigation and discussion of the contribution of each component of our system in the MT quality. Finally, we present conclusions drawn from our investigations, which are in Section 6.

2. RELATED WORK

In this section, we discuss previous work that relate to our research including building bilingual corpora from comparable data, phrase pivot translation, and improving MT on low-resource languages.

2.1. Building Bilingual Corpora

Building parallel corpora from web has been exploited in a long period. One of the first work can be presented in [Resnik 1999]. In order to extract parallel documents from webs, [Li and Liu 2008] used the similarity of the URL and page content. [Utiyama and Isahara 2003] used matching documents to build parallel data. Meanwhile, [Koehn 2005] used manual involvement for building a multilingual parallel corpus. In the work of [Cettolo et al. 2012a], a multilingual corpus was built from subtitles of the TED talks website.

For collecting parallel data from Wikipedia, the task has been investigated in some previous work. In the work of [Kim et al. 2012], parallel sentences are extracted from Wikipedia for the task of multilingual named entity recognition. In [Ştefănescu and Ion 2013], parallel corpora are extracted from Wikipedia for English, German, and Spanish. A recent work proposed by [Chu et al. 2015] extracts parallel sentences before using an SVM classifier to filter the sentences using some features.

For the Asian languages: Japanese, Indonesian, and Malay paired with Vietnamese, there are few bilingual corpora. The Asian Language Treebank, a multilingual parallel corpus of several Asian languages such as Japanese, Indonesian, Malay, Thai, Vietnamese, Myanmar, and English was built manually in [Thu et al. 2016], which is a valuable resource for the languages. Nevertheless, because the corpus is still small with only 20,000 multilingual sentences, and manually building parallel corpora requires many cost of human annotators, automatically extracting large bilingual cor-

[1] https://github.com/nguyenlab/SMT-LowRec

pora becomes an essential task for the development of natural language processing for the languages including cross-language tasks like machine translation. In our work, we built bilingual corpora to improve MT on the Asian low-resource languages. The corpus was built based on Wikipedia's parallel articles that were collected from the articles' title and inter-language link records. Parallel sentences were extracted based on the powerful sentence alignment algorithm [Moore 2002]. The corpus was utilized for improving machine translation on the Asian low-resource languages, in which there has been no work investigated on this task to our best knowledge.

2.2. Pivot Methods

There are three main approaches in pivot methods: cascade, synthetic, and phrase pivot translation (or triangulation). In cascade approaches, source sentences are translated to pivot languages using source-pivot bilingual corpora. Then, the translated pivot sentences are translated to target languages based on pivot-target bilingual corpora ([De Gispert and Marino 2006; Utiyama and Isahara 2007]). In the synthetic approach [De Gispert and Marino 2006], source-pivot or pivot-target translation models are used to generate a synthetic source-target bilingual corpus. For instance, the pivot side of the source-pivot bilingual corpus is translated into the target language using the pivot-target translation model. In the triangulation approach [Cohn and Lapata 2007; Utiyama and Isahara 2007; Wu and Wang 2007], source-pivot and pivot-target bilingual corpora are used to train source-pivot and pivot-target phrase tables. Then, source and target phrases are connected via common pivot phrases by multiplying scores of the source-pivot and pivot-target phrase pairs in the phrase tables. The triangulation approach has shown to be the most effective in pivot methods ([El Kholy et al. 2013; Utiyama and Isahara 2007]).

Pivot translation has been applied in some previous work. [Cettolo et al. 2011] applied pivot methods for Arabic-Italian translation via English and showed the effectiveness. In [Marujo et al. 2011], pivot methods were used in translation from Brazilian Portuguese texts to European Portuguese. For a large-scale data set, [Koehn et al. 2009] applied pivot methods on the multilingual Acquis corpus.

Although the phrase pivot translation method has been applied successfully in several language pairs, there is no prior work for investigating whether this method can be effective on the Asian low-resource languages of Japanese, Indonesian, Malay, and Vietnamese. Meanwhile, for our goal in this research is to improve MT on the low-resource languages, we investigated the phrase pivot translation method on the Asian low-resource languages, which take advantage of existing bilingual corpora for this task.

2.3. Phrase-based Machine Translation on Low-Resource Languages

The topic of improving SMT for low-resource languages has been investigated in some previous work. [Jeff et al. 2011] proposed a word clustering algorithm using monolingual data to improve word alignment for SMT. In [Irvine 2013], monolingual and comparable data sets were used to improve SMT on low resource conditions based on approaches of translating unknown words, bilingual lexicon induction, and scoring phrase tables. [Musleh et al. 2016] presented a work of data collection from the web to improve SMT for low-resource language pairs. One of the most recent work of SMT for low-resource languages has been presented in [Wang et al. 2016], which introduced approaches to source language adaptation for resource-poor SMT using a large bi-text for a related resource-rich language and obtained significant improvement.

The issue of MT on low-resource languages can be also addressed by leveraging additional resources in separate strategies such as building bilingual corpora from comparable data [Resnik 1999; Utiyama and Isahara 2003; Chu et al. 2014], utilizing mono-

lingual data [Ravi and Knight 2011; Nuhn et al. 2012; Saluja et al. 2014], or pivoting bilingual corpora [De Gispert and Marino 2006; Cohn and Lapata 2007; Utiyama and Isahara 2007; Wu and Wang 2007]. In our work, we first investigated the two strategies separately: building bilingual corpora from comparable data and pivoting methods to take advantage of additional resources. Then, we introduced a combined model, which is based on the robust linear interpolation method for combining translation models [Sennrich 2012], to exploit the potential of the additional resources most effectively to improve MT on the low-resource languages. The model was applied for the Asian low-resource languages of translation from Japanese, Indonesian, Malay to Vietnamese, in which there is no prior work of MT on these language pairs despite the fact that these are widely used in the world.

In our work, one of the main contribution is to deal with the out-of-vocabulary (OOV) problem, in which we do not have large parallel corpora to cover vocabulary for new test sets. The additional resources obtained from bridging via pivot languages or comparable data such as Wikipedia can provide more vocabulary. In the work of [Habash 2008], they propose several techniques to deal with the out-of-vocabulary problem such as: connecting the OOV words with existing in vocabulary words, connecting using a manually created dictionary, or using transliteration and achieved more than +1 BLEU points improvement. A recent work of [Luong et al. 2015] addressed the OOV problem on neural machine translation. In their work, a word dictionary is built using a word aligner. Then, OOV words are annotated so that they are not translated in a neural machine translation model, but in a post-processing step. The model achieved an improvement of +2.8 BLEU points. Both of the work do not make use external data to solve the problem. Another work using external resource to overcome the OOV problem on SMT is reported in [Razmara et al. 2013]. In this work, they used a graph propagation approach to induce lexicon from monolingual corpora. In our work, we utilize parallel corpora from different methods of pivoting and building bilingual corpora and then combine the new extracted information to solve the OOV problem.

3. METHODS

We present methods used in this research to improve MT for several low-resource Asian languages of translation from Japanese, Indonesian, and Malay to Vietnamese based on two strategies: building bilingual corpora and exploiting existing bilingual corpora. For the first strategy, we built bilingual corpora for the Asian language pairs from comparable data using Wikipedia parallel titles. For the second strategy, existing bilingual corpora of source and target languages paired with a common language (e.g. English) were exploited using phrase pivot translation. In order to create an effective model to further improve MT on low-resource languages, a combined model was introduced, which utilizes the additional resources extracted from the two strategies.

3.1. Building Bilingual Corpora

Bilingual corpora are essential resources for training MT models. Nevertheless, large bilingual corpora are unavailable for most language pairs, especially for the low-resource languages. Therefore, building bilingual corpora is an important task to improve MT on the low-resource languages. We exploited Wikipedia parallel titles to build bilingual corpora for Japanese, Indonesian, and Malay paired with Vietnamese.

Wikipedia is a large resource that contains a number of articles in many languages in the world. The freely accessible resource is a kind of comparable data in which many articles are in the same domain in different languages. We can exploit this resource to build bilingual corpora, especially for low-resource language pairs. In order to build a bilingual corpus from Wikipedia, we first extracted parallel titles of Wikipedia articles. Then, pairs of articles were collected based on the parallel titles. Finally, sentences in

the article pairs were aligned to extract parallel sentences. We describe the procedure of building the corpus in more detail in this section.

3.1.1. Extracting Parallel Titles. In order to extract parallel titles of Wikipedia articles, we used two resources for each language from the Wikipedia database dumps:[2] the articles' titles and IDs in a particular language (the file ends with *-page.sql.gz*) and the interlanguage link records (the file ends with *-langlinks.sql.gz*).

Table I. **Wikipedia database dumps' resources for extracting parallel titles**; **page (KB)**: the size of the articles' IDs and their titles in the language; **langlinks (KB)**: the size of the interlanguage link records; we used the database *updated on 2017-01-20*.

No.	Data	page (KB)	langlinks (KB)
1	Japanese	114,382	112,722
2	Indonesian	57,921	72,117
3	Malay	21,791	56,173
4	Vietnamese	92,541	111,420

We aim to build bilingual corpora for several Asian low-resource language pairs including Japanese, Indonesian, and Malay paired with Vietnamese, in which there are few existing bilingual corpora. Therefore, we collected the Wikipedia database dumps of the languages: Japanese, Indonesian, Malay, and Vietnamese. Table I presents the Wikipedia database dumps that we used to extract parallel titles.

3.1.2. Collecting and Preprocessing Parallel Articles. After parallel titles of Wikipedia articles were extracted, we collected the article pairs using the parallel titles. We implemented a Java crawler for collecting the articles. The collected data set was then carefully processed in hierarchical steps from articles to sentences, then to word levels. First, noisy characters were removed from the articles. Then, for each article, sentences in paragraphs were splitted so that there is one sentence per line. For each sentence, words were tokenized that separated them from punctuations. We used Moses scripts[3] for word tokenization on Indonesian, Malay, Vietnamese, and we used the Mecab tokenizer[4] for Japanese. Statistics on the extracted parallel titles and collected article pairs are presented in Table II.

Table II. **Extracted and processed data from parallel titles**; **ja**: Japanese, **id**: Indonesian, **ms**: Malay, **vi**: Vietnamese; **Crawled Src Art. (Crawled Trg Art.)**: the number of crawled source (target) articles using the title pairs for each language pair; **Art. Pairs**: the number of parallel articles processed after crawling; **Src Sent. (Trg Sent.)**: the number of source (target) sentences in the article pairs after preprocessing (removing noisy characters, empty lines, sentence splitting, word tokenization).

No.	Data	Title pairs	Crawled Src Art.	Crawled Trg Art.	Art. Pairs	Src Sent.	Trg Sent.
1	ja-vi	103,816	92,120	87,436	82,483	2,358,926	1,864,782
2	id-vi	159,247	149,974	128,530	121,673	1,201,848	1,878,855
3	ms-vi	133,651	118,647	116,620	105,692	560,042	1,256,468

[2]https://dumps.wikimedia.org/backup-index.html
[3]https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer
[4]https://taku910.github.io/mecab/

3.1.3. Aligning Parallel Sentences. Sentence alignment is an essential task in building parallel corpora. Among the three main approaches in sentence alignment: length-based which is based on the number of words or characters [Brown et al. 1991; Gale and Church 1993], word-based which is based on word correspondences [Kay and Röscheisen 1993; Chen 1993; Wu 1994; Melamed 1996; Ma 2006], and the hybrid of length-based and word-based [Moore 2002; Varga et al. 2007], the hybrid method of [Moore 2002] achieved the best performance as the evaluation of [Singh and Husain 2005]. In our work, for each parallel article pair, we aligned sentences using the hybrid method [Moore 2002]. There are several reasons to adopt the hybrid method for aligning parallel sentences in this task. First, the length-based method has been applied successfully in close languages such as English-French; however, the Asian low-resource languages including Japanese, Indonesian, Malay, and Vietnamese are not close languages excluding Indonesian-Malay. Second, the Wikipedia bilingual articles vary greatly in terms of the number of sentences in bilingual articles and the number of words in sentence pairs. Therefore, we adapted the hybrid method that combines the length-based and word correspondences to extract the parallel corpus.

Table III. **Extracted parallel corpora**; **Pairs**: the number of parallel sentences; **Src Words (Trg Words)**: the number of source (target) words; **Src Vocab. (Trg Vocab.)**: the source (target) vocabulary size; **Src Len (Trg Len)**: the average length of source (target) sentences

No.	Data	Pairs	Src Words	Trg Words	Src Vocab.	Trg Vocab.	Src Len	Trg Len
1	ja-vi	18,684	357,347	294,434	37,203	24,682	19	16
2	ms-vi	48,146	696,403	807,268	55,075	46,041	14	17
3	id-vi	72,132	946,971	1,061,452	65,263	53,692	13	15

3.1.4. Extracted Wikipedia Corpus. We obtained parallel corpora of the Asian low-resource language pairs as described in Table III.

These corpora represent of our main contributions (as mentioned in the introduction). To the best of our knowledge, this is one of the first efforts in building parallel corpora for these language pairs. As such, they constitute a valuable resource, which can be re-used in future studies.

3.2. Pivot Methods

The previous section presented our strategy of building bilingual corpora to augment the training data for MT models. Besides this strategy, we can take advantage of existing bilingual corpora via intermediate language(s) called pivot methods if the bilingual corpora between source and target languages paired with the pivot language are available. Specifically, in order to translate from the source language to the target language, pivot language(s) can be used as a bridge for translations if there exist source-pivot and pivot-target bilingual corpora.

In various approaches of pivot methods, the *triangulation* (or phrase pivot translation) has been shown to be the best approach [Utiyama and Isahara 2007]. In triangulation [Cohn and Lapata 2007; Utiyama and Isahara 2007; Wu and Wang 2007], source-pivot and pivot-target bilingual corpora are used to train phrase tables. Then, the source and target phrases are connected via common pivot phrases.

Given a source phrase s and target phrase t of the source-pivot phrase table T_{SP_s} and the pivot-target phrase table T_{P_tT}, the phrase translation probability is estimated via common pivot phrases p based on the following feature function [Wu and Wang 2007].

$$\phi(t|s) = \sum_{p \in (T_{SP_s}) \cap (T_{P_tT})} \phi(p|s)\phi(t|p) \quad (1)$$

Previous work showed the effectiveness of this method when the source-target bilingual corpus (direct corpus) is unavailable or in limited size. In this work, we utilized the phrase pivot translation approach to take advantage available bilingual corpora of the Asian low-resource languages paired with a common language (typically English) to improve MT on the language pairs.

3.3. A Combined Model for Phrase-based Machine Translation on Low-Resource Languages

In previous two sections, we present two strategies in leveraging additional resources to improve MT on the Asian low-resource language pairs: building bilingual corpora using Wikipedia parallel titles and exploiting existing bilingual corpora based on the phrase pivot translation. The question is how can we utilize the different additional resources of the extracted Wikipedia corpus and the pivoted phrase table to effectively improve MT on the low-resource languages. In this section, we discuss the method of linear interpolation in combining translation models, and a robust technique for this task based on perplexity optimization. Then, we introduce a combined model that adapts the linear interpolation based on perplexity optimization, which combines the additional resources to build an effective model for MT on the low-resource languages.

3.3.1. Translation Model Combination.

Linear Interpolation. In order to combine different translation models, a well-known approach can be utilized called linear interpolation, which is formulated as follows.

$$p(x|y;\lambda) = \sum_{i=1}^{n} \lambda_i p_i(x|y) \tag{2}$$

where λ_i is the interpolation weight of each model i, and with $\sum_i \lambda_i = 1$.

One of the main factors in the approach is the interpolation weights λ. The approach has been applied in many systems using different techniques of setting the interpolation weights such as: uniform weights [Cohn and Lapata 2007], using different interpolation coefficients [Axelrod et al. 2011], and monolingual metrics to set the weights [Koehn et al. 2010]. Nevertheless, the techniques have been shown that they often do not perform optimally as reported in [Yasuda et al. 2008]. Alternatively, a robust method based on perplexity optimization for linear interpolation has been proposed in [Sennrich 2012], which showed advantages in combining translation models that we discuss in more detail below.

Perplexity Minimization. The method of linear interpolation based on perplexity optimization [Sennrich 2012] has been proposed in the task of combining different translation models in domain adaptation, in which the interpolation weights were searched that minimize the cross-entropy between the interpolation model and a tuning set of word/phrase alignments. As discussed in previous work, the method has been shown effectiveness in adapting translation models to translationese and achieved statistically significant improvements in translation quality [Lembersky et al. 2013]. This is because perplexity is a good differentiator between original and translated texts, and the perplexity is optimized based on development sets.

In the perplexity minimization method for linear interpolation of combining translation models, our goal is to determine the interpolation weights λ_i based on a development corpus. The cross-entropy $H(p)$ and the perplexity are equivalent in the purpose of optimization, which is defined as follows.

$$H(p) = -\sum_{x,y} \tilde{p}(x,y) log_2 p(x|y) \tag{3}$$

The goal is to minimize the perplexity or reduce the cross-entropy, in which the model is better to predict the development set. Given a development corpus, a phrase table is trained using the configuration as in training the translation models. From the development phrase table, we obtained phrase pairs x, y and their empirical probabilities $\tilde{p}(x, y)$. The objective function of minimization of the cross-entropy is described as:

$$\hat{\lambda} = argmin_{\lambda} - \sum_{x,y} \tilde{p}(x,y) log_2 p(x|y;\lambda) = argmin_{\lambda} - \sum_{x,y} \tilde{p}(x,y) log_2 \sum_{i=1}^{n} \lambda_i p_i(x|y) \quad (4)$$

where the weight vector λ is the argument, and p is the probability of the interpolation model.

3.3.2. A Combined Model for SMT on Low-Resource Languages. As discussed in the previous sections, we focus on two strategies: building a bilingual corpora and exploiting existing bilingual corpora using the phrase pivot translation. In the first strategy, after a bilingual corpus was built, we trained a phrase table using a standard phrase-based configuration. In the second strategy, a source-target phrase table was generated using the triangulation. Additionally, we utilized existing direct bilingual corpora of the source and target languages to train a direct phrase table. The three phrase tables were combined using the linear interpolation based on perplexity minimization method to generate a new phrase table that combines the different additional resources. The combined model is described as:

$$p(t|s) = \lambda_d p_d(t|s) + \lambda_a p_a(t|s) + \lambda_{tr} p_{tr}(t|s) \quad (5)$$

where $p_d(t|s)$, $p_a(t|s)$ and $p_{tr}(t|s)$ stand for the translation probabilities of the *direct*, *align*, and *triangulation* phrase tables, respectively; interpolation parameters: λ_d, λ_a, and λ_{tr} (where $\lambda_d + \lambda_a + \lambda_{tr} = 1$). The lambda values are automatically tuned given a development set using the script.[5]

We named the components in this model: *direct* (for the phrase table trained on the direct bilingual corpus), *align* (for the phrase table trained on our extract Wikipedia corpus), and *triangulation* (for the triangulated phrase table). We evaluated the effectiveness of the combined model in improving MT on the Asian low-resource languages, which is presented in the section of experiments.

4. EXPERIMENTS

In this section, we present experiments of improving MT on the Asian low-resource languages: translation from Japanese, Indonesian, and Malay to Vietnamese, in which there is no prior work of MT on these language pairs. First, we conducted experiments on several available direct bilingual corpora using a standard phrase-based machine translation configuration. Then, the bilingual corpora built from Wikipedia bilingual titles were utilized to improve the MT on the low-resource languages. Existing bilingual corpora between the Asian languages paired with English were exploited using the phrase pivot translation to enhance MT on the Asian language pairs. Finally, we present experimental results of the combined model (Section 3.3.2), which adapted the linear interpolation based on perplexity optimization to combine the additional resources for the goal of creating an effective solution for improving MT on the Asian low-resource languages.

[5]https://github.com/moses-smt/mosesdecoder/tree/master/contrib/tmcombine

4.1. Setup

4.1.1. Training Data. We describe the data sets used in our experiments including: the bilingual training data (direct bilingual corpora, bilingual corpora for training phrase pivot translation), the development and test sets, and the monolingual data for training language model.

Parallel data. For Japanese-Vietnamese, we used a small bilingual data extracted from the TED talks [Cettolo et al. 2012b],[6] and the Bible data.[7] For Malay-Vietnamese and Indonesian-Vietnamese, we used the TED data [Cettolo et al. 2012b]. First, the monolingual data of the TED talks were collected from the web[8] for Malay, Indonesian, and Vietnamese. Then, the Malay-Vietnamese and Indonesian-Vietnamese parallel sentences were aligned based on the *talk id* and the *seekvideo id* from the collected data. The bilingual training data sets are described in Table IV.

Table IV. **Direct Bilingual Corpora**; **ja**: Japanese, **vi**: Vietnamese, **ms**: Malay, **id**: Indonesian; **Src Words, Trg Words**: number of source, target words; **Src Vocab, Trg Vocab**: the source, target vocabulary; **Src Avg len, Trg Avg len**: the average length of source, target sentences

Data set	Sentence Pairs	Src Words	Trg Words	Src Vocab	Trg Vocab	Src Avg len	Trg Avg len
ja-vi	83,313	2,076,083	2,138,623	37,689	19,411	25	25
ms-vi	17,655	133,528	191,229	10,140	5,664	8	11
id-vi	210,495	1,564,314	2,282,637	40,097	21,025	7	11

For pivot translation, we exploited the existing bilingual corpora of the Asian languages paired with English, which typically exist because of the popularity of English. For Japanese-Vietnamese experiments, we used the Japananese Kyoto corpus [Neubig 2011].[9] The bilingual data of English-Vietnamese includes the VLSP corpus,[10] the English-Vietnamese training data in the IWSLT 2015,[11] and an in-house bilingual corpus used in the system [Trieu et al. 2015] participated in the IWSLT 2015. For Indonesian-Vietnamese and Malay-Vietnamese experiments, we utilized the TED talks data for Indonesian-English, Malay-English, and English-Vietnamese as the procedure that we extracted the direct corpus. The bilingual corpora for pivot translation are described in Table V.

Table V. **Bilingual corpora for phrase pivot translation**; **en**: (English) was used for pivot

Data set	Sentence Pairs	Src Words	Trg Words	Src Vocab	Trg Vocab	Src Avg len	Trg Avg len
ja-en	329,882	6,085,131	5,911,486	114,284	161,655	18	18
ms-en	15,635	118,602	135,848	9,788	10,401	8	9
id-en	228,886	1,706,053	1,989,335	42,651	44,712	7	9
en-vi	377,736	4,446,502	3,562,696	36,661	67,325	12	9

[6] https://wit3.fbk.eu/mt.php?release=2012-02-plain
[7] http://homepages.inf.ed.ac.uk/s0787820/bible/
[8] https://wit3.fbk.eu/
[9] http://www.phontron.com/kftt/
[10] http://vlsp.hpda.vn:8080/demo/?page=resources
[11] http://workshop2015.iwslt.org/

Development and Test Sets. For development and test data sets, we utilized the Asian Language Treebank corpus [Thu et al. 2016], a multilingual corpus of Asian languages such as Japanese, Indonesian, Malay, Vietnamese, Thai, etc and English including 20k multilingual sentences. We randomly extracted 2k sentences for development and 2k sentences for test sets of Japanese, Indonesian, and Malay paired with English. The development sets were also used as the development corpus for the combined model, which adapted the linear interpolation based on perplexity minimization. Tables VI and VII present the development and test sets.

Table VI. **Development sets**

Data set	Sentence Pairs	Src Words	Trg Words	Src Vocab	Trg Vocab	Src Avg len	Trg Avg len
ja-vi	2,000	68,681	69,368	7,564	5,129	34	35
ms-vi	2,000	47,647	69,368	7,608	5,129	24	35
id-vi	2,000	47,638	69,368	7,956	5,129	24	35

Table VII. **Test sets**

Data set	Sentence Pairs	Src Words	Trg Words	Src Vocab	Trg Vocab	Src Avg len	Trg Avg len
ja-vi	2,084	67,012	68,917	7,890	5,384	32	33
ms-vi	2,084	48,867	68,917	7,792	5,384	23	33
id-vi	2,084	48,159	68,917	8,130	5,384	23	33

Monolingual Data. We employed a large monolingual data set of Vietnamese for training language model including 16M sentences extracted from articles of the website *baomoi.com*[12] as used in our IWSLT 2015 systems [Trieu et al. 2015]. We present the monolingual data in Table VIII.

Table VIII. **Monolingual data set**; **Words**: number of words, **Vocab**: vocabulary, **Avg len**: the average length of sentences

Sentences	Words	Vocab	Avg len
16,201,114	485,087,517	850,650	30

All data sets were preprocessed for sentence split, word tokenization, and truecase using the Moses script.[13]

4.1.2. Training Details. We describe configurations for training phrase-based MT, building bilingual corpora (Section 3.1), and phrase pivot translation (Section 3.2).

Phrase-based SMT. We conducted experiments of phrase-based MT using the well-known Moses toolkit [Koehn et al. 2007]. Word alignment was trained using the GIZA++ [Och and Ney 2003] with the configuration *grow-diag-final-and*. For language model, we trained a 5-gram language model on the large Vietnamese monolingual data set using the KenLM [Heafield 2011]. Parameters were tuned using the batch MIRA [Cherry and Foster 2012]. The evaluation was performed based on the typically used BLEU [Papineni et al. 2002].

[12]http://www.baomoi.com/
[13]https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer

Building Bilingual Corpora. In building bilingual corpora from Wikipedia parallel titles, we implemented the powerful algorithm Microsoft sentence aligner [Moore 2002] using Java. The method includes three phases: length-based phase (aligning sentences by length), word alignment (training a word alignment model using sentence pairs extracted from the length-based phase), and length-and-word-based phase (combining the word alignment with the length-based). We set the threshold of the length-based phase as 0.99, and the threshold of the length-and-word-based phase 0.9 to ensure the high accuracy.

Phrase Pivot Translation. For the phrase pivot translation, we implemented the well-known triangulation method [Wu and Wang 2007] using Java. As discussed in [El Kholy et al. 2013], a big issue of the triangulation method is a triangulated phrase table is generated with very big size and contains noisy phrase pairs. In our implementation, we used a $n - best$ filtering technique in which a set of n best target phrases were filtered given each source phrase (n was set as 10 in our experiments).

4.2. Direct Systems

The main problem of MT on low-resource languages is unavailable large bilingual data. For the case of the Asian low-resource languages, we take advantage of several available small direct bilingual corpora to train phrase-based MT models (*direct systems*), which require improvement from the additional resources. The direct bilingual data sets (Table IV) were used to train direct systems using the *Phrase-based SMT configuration* as we described. Then the systems were evaluated using the development and test sets (Tables VI and VII).

Table IX. **Experimental results of direct systems** (BLEU); phrase-based MT systems trained on the direct bilingual corpora; **Dev, Test**: evaluated on the development and test sets

System	Dev	Test
Japanese-Vietnamese	9.91	10.80
Malay-Vietnamese	15.52	17.80
Indonesian-Vietnamese	25.76	26.80

We obtained the results of the direct systems as shown in Table IX. From the experimental results, the Indonesian-Vietnamese and Malay-Vietnamese were somewhat reasonable (26.80 and 17.80 BLEU points). However, it was a low performance on the Japanese-Vietnamese (10.80 BLEU point). We showed how our models can improve the performance in the next sections.

4.3. Results on Building Bilingual Corpora

As presented in Section 3.1, we built parallel corpora based on Wikipedia parallel titles for the Asian low-resource languages. The corpora were used to train MT models. Then, we evaluated the performance of the models trained from the extracted Wikipedia corpora on the development and test sets. The experimental results are presented in Table X.

Using our extracted corpus for training MT models, we obtained promising results such as: Indonesian-Vietnamese (24.65 BLEU), Malay-Vietnamese (21.00 BLEU), which show that although the corpora are built based on a resource with noisy bilingual texts, we can still obtain the performance comparable with the direct systems, which were trained on the manually translated texts. The low scores obtained for

Table X. **Experimental results using our extracted Wikipedia corpora** (BLEU); **Aligned**: the number of parallel sentences in the corpus

System	Aligned	Dev	Test
Japanese-Vietnamese	18,684	4.90	5.82
Malay-Vietnamese	48,146	18.45	21.00
Indonesian-Vietnamese	72,132	23.86	24.65

Japanese-Vietnamese is due to the small number of parallel sentences for this language pair.

4.4. Results on Phrase Pivot Translation

In exploiting existing bilingual corpora using the phrase pivot translation, promising results were obtained, which nearly approach the performance of the direct systems. Table XI presents the experimental results using phrase pivot translation. The results show that the additional resources exploited based on the pivot method are effective for MT on the Asian low-resource languages.

Table XI. **Experimental results of pivot translation** (BLEU);

System	Dev	Test
Japanese-Vietnamese	8.44	9.40
Malay-Vietnamese	16.33	18.71
Indonesian-Vietnamese	25.99	26.71

The pivot translation showed promising results when the performance was approximate to the direct systems' performance (9.40 vs 10.80 BLEU points on Japanese-Vietnamese, 18.71 vs. 17.80 BLEU points on Malay-Vietnamese, and 26.71 vs. 26.80 BLEU points on Indonesian-Vietnamese). These results show that pivot translation can be useful in addressing the problem posed by the unavailability of direct bilingual data. In comparison among language pairs, the Japanese-Vietnamese still obtained the lower performance than the two other language pairs, which is also similar to the case of the direct systems as we discussed in Section 4.2.

4.5. Results on Combination Models

Previous sections show the promising results in both strategies of using our extracted Wikipedia corpus and utilizing existing bilingual corpora via phrase pivot translation. Our goal is to build systems which are able to exploit the additional resources most effectively to further improve MT on the low-resource languages. Our combined model, which combined the direct data and the additional resources (the extracted corpus from Wikipedia (*align*) and the pivot translation (*pivot*)), was used for this goal. We present experimental results of the combined model with various combination between the translation models in Tables XII-XIV. To assess whether the gains in performance of the combined models are pertinent, we conducted statistical significance test with 95% confidence interval using the paired bootstrap resampling [Koehn 2004] (the statistically significant improvements are marked * in the result tables).

For Japanese-Vietnamese, the combined model achieved significant improvement compared with the direct model as described in Table XII. While the align component improved +0.4 BLEU because of the small extracted Wikipedia corpus, the pivot component obtained the higher improvement (+2.03). The combined system of all three components achieved the significant improvement, which can be an effective solution for improving MT on Japanese-Vietnamese, a language pair with big challenges in

Table XII. **Experimental results of the combined model on Japanese-Vietnamese** (BLEU); **direct, pivot, align**: the system trained on the direct bilingual corpus, using the phrase pivot translation, and our extracted Wikipedia corpus, respectively; **direct-align, direct-pivot, direct-pivot-align**: various combinations between the translation models using the linear interpolation with perplexity minimization;

System	Dev	Test
direct	9.91	10.80
pivot	8.44	9.40
align	4.90	5.82
direct-align	10.10	11.21 (+0.41)
direct-pivot	11.57	12.83 (+2.03)*
direct-pivot-align	11.71	12.88 (+2.08)*

MT because of not only differences in language structures but the limited amount of bilingual training data.

Table XIII. **Experimental results of the combined model on Malay-Vietnamese** (BLEU);

System	Dev	Test
direct	15.52	17.80
pivot	16.33	18.71
align	18.45	21.00
direct-align	21.17	23.91 (+6.11)*
direct-pivot	17.92	20.61 (+2.81)*
direct-pivot-align	22.37	25.00 (+7.20)*

For Malay-Vietnamese, since using our extracted Wikipedia data and using phrase pivot translation obtained promising results, and the direct system was trained on a small data, the combined model shows a significant improvement up to +7.2 BLEU, in which our extracted corpus shows the effectiveness with +6.11 BLEU. Table XIII presents the Malay-Vietnamese experimental results.

Table XIV. **Experimental results of the combined model on Indonesian-Vietnamese** (BLEU);

System	Dev	Test
direct	25.76	26.80
pivot	25.99	26.71
align	23.86	24.65
direct-align	27.49	28.42 (+1.62)*
direct-pivot	27.24	27.97 (+1.17)*
direct-pivot-align	28.13	28.86 (+2.06)*

For Indonesian-Vietnamese, although the direct system was trained on a larger direct bilingual data and obtained a much higher performance compared to the experiments on the other language pairs, the combined model still achieved the significant improvement in both combining the direct with the additional resources separately, or combining all components, as described in Table XIV. Our extracted corpus, with the larger size compared to the corpus in other language pairs, achieved the significant improvement of +1.62 BLEU, whereas the pivot component also showed the effectiveness with +1.17 BLEU. As results in Japanese-Vietnamese and Malay-Vietnamese, the combined system of the three components achieved the best performance, which confirms

the effectiveness of the combined model in improving MT on the Asian low-resource languages.

5. ANALYSIS

In this section, we analyzed the contribution of each additional resource in improving MT on the Asian low-resource languages by further investigating the aspect of out of vocabulary (OOV), which shows the ratio of vocabulary that was not translated and covered by direct systems. For MT systems trained on low-resource bilingual data, the OOV problem typically happens in such data sparseness, which leads to a low MT performance. Therefore, our work is to leverage additional resources in dealing with the problem to improve MT on the languages. First, we analyze the OOV ratio of direct systems and the reduction of OOV ratio by leveraging the additional resources. Then we compare the correlation between the OOV ratio and the improvement on the different language pairs. Additionally, several sample translations were illustrated, which show the improvement generated from our combined model of the additional resources. Finally, we discuss several limitations of this research, which need to be addressed, and some directions are opened for further improvement in future work.

5.1. Dealing with the Out of Vocabulary Problem

The statistics of the OOV ratio are presented in Table XV. The OOV ratios of the pivot and align models were not much higher than the ratios of the direct models. For the align systems, which were trained on our extracted Wikipedia corpus, the larger the corpus size obtained the higher performance (35.36% of Malay-Vietnamese and 27.32% of Indonesian-Vietnamese versus 47.82% of Japanese-Vietnamese).

Table XV. **Statistics of the out of vocabulary ratio** (%)

System	ja-vi	ms-vi	id-vi
direct	29.87	43.06	19.56
pivot	42.76	44.57	18.49
align	47.82	35.36	27.32
direct-pivot	23.22	39.22	16.77
direct-align	25.06	28.55	16.05
direct-pivot-align	21.09	26.94	14.42

The OOV ratios in the combined systems were lower than the ratios in the direct systems and also in the align and pivot systems as presented in Table XV. Additionally, the combined systems show more contribution when the direct systems got a high OOV ratio. This indicates that our combined model becomes more effective for MT on fewer bilingual data.

5.2. Bilingual Data, BLEU, and OOV

Leveraging additional resources can help to improve the performance when the additional resources contain informative vocabulary. However, the question is how each of the additional resources contributes to the improvement, and whether the strategy of leveraging additional resources always helps regardless of language pairs. We analyzed the correlation of the direct bilingual corpus size, the improvement in BLEU, and the reduction of OOV ratio. Additionally, we conducted a comparison between the additional resources as well as between language pairs, which are presented in Table XVI.

Several findings can be drawn from the analyses. First, the more the OOV reduction, the more the improvement. In all cases, the combined systems of **B-P-A** (direct-pivot-align) showed the highest reduction in OOV, and also the highest improvement

Table XVI. **Correlation of improvement and reduction of OOV ratio**; **direct, s-p, p-t, wiki**: the corpus size of the direct, source-pivot and pivot-target in pivot translation, and our extracted Wikipedia corpus, respectively; **B, B-P, B-A, B-P-A**: the direct system, and the combined systems of direct-pivot, direct-align, direct-pivot-align; there are two lines for each system: the **BLEU score** (on the first line) and the **[OOV ratio]** (on the second line) as well as the increase in BLEU and the reduction in OOV ratio generated by the combined systems

System	direct	s-p	p-t	wiki	B	B-P	B-A	B-P-A
ja-vi	83k	329k	456k	18k	10.80 [29.87]	12.83 (+2.03) [23.22 (-6.65)]	11.21 (+0.41) [25.06 (-4.81)]	12.88 (+2.08) [21.09 (-8.78)]
id-vi	210k	228k	377k	72k	26.80 [19.56]	27.97 (+1.17) [16.77 (-2.79)]	28.42 (+1.62) [16.05 (-3.51)]	28.86 (+2.06) [14.42 (-5.14)]
ms-vi	17k	15k	377k	48k	17.80 [43.06]	20.61(+2.81) [39.22 (-3.84)]	23.91 (+6.11) [28.55 (-14.51)]	25.00 (+7.20) [26.94 (-16.12)]

in BLEU (e.g, Japanese-Vietnamese: -8.78 OOV and +2.08 BLEU, Malay-Vietnamese: -16.12 OOV and +7.20 BLEU). Second, the improvement of the align systems depend on the direct bilingual corpus size for training the direct systems. When the direct bilingual corpus, which is typically in a high quality in terms of alignment, is large, the contribution from the align systems, which were trained on the corpus extracted automatically from comparable data, is less than in the case of small direct bilingual data (e.g, Indonesian-Vietnamese: 210k direct corpus and +1.62 BLEU from the combined direct-align system, Japanese-Vietnamese: 83k direct corpus and +0.41 BLEU contributed from the alignment data). Third, in utilizing bilingual corpora using pivot translation, the pivot systems always show improvement regardless of the language pairs and different domain of the pivot data sets. For the case of Japanese-Vietnamese, we used the TED talks and Bible data sets for training the direct system; we then used the Kyoto corpus for Japanese-English and TED talks data for English-Vietnamese in pivot translation. For Indonesian-Vietnamese and Malay-Vietnamese, all bilingual corpora of Indonesian-Vietnamese, Malay-Vietnamese, Indonesian-English, Malay-English, and English-Vietnamese were extracted from the same data: the monolingual data of TED talks for these four languages. Fourth, in comparison between the additional components (pivot and align) to the combined model, the pivot component contributed an improvement from +1 to +2 BLEU points (Japanese-Vietnamese: +2.03, Indonesian-Vietnamese: +1.17, Malay-Vietnamese: +2.81); meanwhile, the contribution of the alignment data is usually higher than the pivot model (+1.62 vs. +1.17 in Indonesian-Vietnamese, +6.11 vs. +2.81 in Malay-Vietnamese). Fifth, we seek to an analysis of the effects of leveraging additional resources on the language pairs. The finding is that additional resources do not always improve MT performance, and the size of direct bilingual corpus is not always the main issue. When comparing the direct corpus size and the improvement from the additional resources between the two language pairs Japanese-Vietnamese and Malay-Vietnamese, even though a larger direct bilingual data of the Japanese-Vietnamese was used (83k vs. 17k), and the OOV ratios are lower than that of Malay-Vietnamese (29.87 vs. 43.06 (direct), 21.09 vs. 26.94 (direct-pivot-align)), the performance of the Japanese-Vietnamese was still lower than that of Malay-Vietnamese (10.80 vs. 17.80, 12.88 vs. 25.00). Even when translation models covered a higher ratio of vocabulary, another main challenge of MT also relates to how the words and phrases can be ordered in the translation output. The issue becomes potential in this case when Malay and Vietnamese share the same language structure of Subject-Verb-Object whereas Japanese is different with Subject-Object-Verb. For further improvement of MT on the Asian low-resource languages, an investigation of reordering is also need to be considered in addition to leveraging additional resources although the strategy has been utilized effectively with promising results.

5.3. Sample Translations

We illustrate several sample translations in Table XVII, which show the OOV problem in the direct systems as well as the contribution of the additional resources in the combined systems to deal with the OOV problem.

Table XVII. **Sample Translations**; for each sample, there are components: **source** (input sentence of ja(Japanese), id (Indonesian), and ms (Malay); **reference** (reference sentences of vi (Vietnamese)); **meaning**: the English meaning translated by the authors in *the italic*; **direct**: the translation generated by the direct systems; **direct-pivot-align**: the translation generated by the combined model. OOV words are indicated by the [square brackets]. The **bold** indicates the correct translations generated by the the combined systems

System	Sample Translations
	Japanese-Vietnamese
source (ja)	ソマリア の スポーツ 担当 大臣 も 、 入院 し て 、 重態 に ある こと が 報告 さ れ て いる 。
reference (vi)	bộ trưởng Thể thao Somali cũng được báo cáo là đã **nhập viện** và trong tình trạng nguy kịch .
meaning	*the Somali minister for sports is also reported **to be hospitalized** and in critical condition .*
direct	Bộ trưởng thể thao của Somalia , rồi nhập vào [重態] , được báo cáo là có thể .
direct-pivot-align	Bộ trưởng phụ trách thể thao của Somalia , cũng đã được **nhập viện** , hy vọng sẽ có được báo cáo .
	Indonesian-Vietnamese
source (id)	peluncuran yang sukses akan membuat Korea Selatan bisa menjadi pemain di dalam bisnis komersial luar angkasa yang nilai industrinya berkisar USD250 miliar .
reference (vi)	một vụ phóng thành công sẽ có thể giúp Hàn Quốc trở thành một thành viên trong các thương vụ phóng không gian thương mại , một ngành **công nghiệp** trị giá khoảng US $ 250 tỷ .
meaning	*A successful launch will enable Korea to become a member of the commercial space launch , an **industry** worth about US $ 250 billion*
direct	phóng thành công sẽ tạo ra Hàn Quốc có thể trở thành một vận động viên trong kinh doanh thương mại không gian mà giá trị [industrinya] xoay USD250 tỷ .
direct-pivot-align	phóng thành công sẽ tạo ra Hàn Quốc có thể trở thành cầu thủ trong kinh doanh thương mại không gian giá trị **công nghiệp** kéo dài USD250 tỷ .
	Malay-Vietnamese
source (ms)	kenyataan jawatankuasa itu mendapat sokongan daripada pencinta alam sekitar dan juga dari beberapa industri sektor tenaga .
reference (vi)	tuyên bố của Ủy ban đã thu hút được sự ủng hộ từ **các nhà hoạt động môi trường** cũng như từ một số ngành công nghiệp về lĩnh vực năng lượng .
meaning	*The Commission's statement has garnered support from **environmentalists** as well as from some industries in the field of energy .*
direct	tuyên bố của ủy ban này đã nhận được sự ủng hộ từ [pencinta] môi trường và cũng từ một số ngành công nghiệp của khu vực năng lượng .
direct-pivot-align	tuyên bố của ủy ban này đã nhận được sự ủng hộ từ **các nhà môi trường** và cũng từ một số ngành công nghiệp khu vực năng lượng .

From the sample translations, several phrases which were not translated by the direct system (OOV) were translated correctly by our combined system, which show closer meaning to the reference and improve the translation quality. However, several other problems can be observed from these sample translation such as the word *sektor* (means *field*) in the Malay sentence was translated into *khu vực* (means *areas*) by our system, which will be evaluated as an incorrect translation and failed when calculating BLEU although the translation of our system and the reference *lĩnh vực* (also means *field* or *areas*) that are in close meaning. This issue needs to be addressed in the task of choosing appropriate meanings.

5.4. Limitations and Future Work

Although leveraging additional resources has shown the effectiveness in improving MT on the Asian low-resource languages, and the combined model achieved significant improvement, which can be used as an effective solution to solve the task, there are still limitations that need to be addressed and obtain further improvement.

— First, although the extracted Wikipedia corpus contributed to improve MT on the low-resource languages, there are two issues remaining. The quality of the corpus needs to be improved by filtering strategies because bilingual Wikipedia articles are not always aligned to each other in terms of content, document length, sentence length, writing styles, which lead to noisy pairs in the extracted corpus.
— Second, although a large number of Wikipedia articles were collected, only a small number of parallel sentences were aligned. This is a challenge task in working on such kind of comparable data. However, a method proposed in [Munteanu and Marcu 2006] may help, in which parallel sub-sentential fragments are extracted from comparable data. We plan to conduct this task in the future work.
— Third, in terms of language pairs, the performance of Japanese-Vietnamese was still quite low, although improving MT on such language pair with small parallel data and different structures faces a big challenge, which is not only in the case of Japanese-Vietnamese. Even though the goal of this research in leveraging additional resources to improve MT on this low-resource language pair has been achieved, a further analysis in terms of language structures and reordering is needed to improve MT on this language pair.
— Fourth, a direction for dealing with OOV problem is that we can take advantage of the semantic similarity from monolingual data. Specifically, a word pair $s - t$ of the bilingual word dictionary in the sentence alignment algorithm [Moore 2002] can be enriched with two new pairs $s - t'$ and $s' - t$ by the semantic similarity word pairs $s - s'$ and $t - t'$ in monolingual data sets. The enriched dictionary can help to cover more informative vocabulary to improve building bilingual corpora. Meanwhile, for phrase pivot translation, one of the issues of the conventional approach [Wu and Wang 2007] is that information is lose when source and target phrases are connected only by common pivot phrases; however, some pivot phrases share the same meaning even when they are not matched to each other (common pivot). In order to overcome the problem, we can use semantic similarity to extract more informative connections between pivot phrases. This direction is currently in our progress with some promising results, and we plan to integrate it for this task in future research.
— In our current model for building bilingual corpora, we used IBM Model 1 in sentence alignment, which is the same setting as in [Moore 2002]. However, using higher IBM models can produce better extracted corpora, and we leave it in future work.

Table XVIII. **Results on Google Translation**; the translation was performed using Google Translation *https://translate.google.com/* in June, 2018.

Language Pair	Dev	Test
Japanese-Vietnamese	21.55	23.94
Malay-Vietnamese	33.92	34.50
Indonesian-Vietnamese	32.57	36.16

— Recently, Google has published a powerful system for translation using deep neural networks [Wu et al. 2016]. We would like to evaluate how the performance of the

low-resource languages on this system. We conducted translation on Google translate service. The results are reported in Table XVIII. The Google translation can show much better performance than our current results. Although the results are higher than our results, the comparison may be not fair in terms of training data. However, the results can show some important points for the future directions such as investigating the translation quality not only the vocabulary coverage. The performance of Japanese-Vietnamese is still low and also in the two other language pairs. In addition, several examples of the Google Translation's output were presented in Table XIX, which indicate some issues. Although the Google translation results almost cover the vocabulary (no OOV problem), the problem of word choice also need to be tackled. As shown in the examples, "the persons who use carrier pigeons" was translated into "the persons who carry pigeons", or the "great tragedy", "unusual practice" (which mean negative things) were translated into the "wonderful". This is also a problem. We intend to improve the translation quality on the low-resource languages in future work.

Table XIX. **Some examples from the Google Translation's output**

	Examples
Input (Japanese)	法定で彼は「軍事委員会は伝書鳩か何かを使っているのか、私は分からない」と英語で述べた。
Reference (in English)	In court he said, "I don't know — are the military commissions using carrier pigeons or what?" speaking in English.
Google translation	Quy phạm pháp luật trong "Ủy ban Quân Đỗ đang sử dụng một cái gì đó hoặc người vận chuyển chim bồ câu, tôi không biết", ông nói bằng tiếng Anh.
Meaning	Legal in "Quan Do committee is using something or person who carries pigeons, I don't know", he said in English.
Input (Indonesian)	"Ini adalah tragedi besar bagi penduduk Pichilemu dan tidak bisa diabaikan."
Reference (in English)	"It's a great tragedy for the Pichileminian and couldn't pass unnoticed."
Google translation	"đây là một thảm kịch tuyệt vời cho cư dân Pichilemu và không thể bỏ qua."
Meaning	this is a wonderful tragedy for the Pichileminian and couldn't pass unnoticed
Input (Malay)	Sehingga kebelankangan ini, pengeboman berani mati adalah satu amalan luar biasa di Somalia.
Reference (in English)	Until recently, suicide bombings were an unusual practice in Somalia.
Google translation	Ở mức độ này, một vụ đánh bom tự sát là một thực tế tuyệt vời ở Somalia.
Meaning	In this level, a suicide bombing was a wonderful practice in Somalia.

6. CONCLUSION

Phrase-based MT systems achieve high quality when trained on large bilingual corpora. Nevertheless, such large bilingual corpora are unavailable for most language pairs in the world, which cause a bottleneck for MT on these low-resource languages. In a more specific case, the Asian languages such as Japanese, Indonesian, Malay, and Vietnamese also belong to that issue despite the fact that the languages are widely used in the world. Furthermore, there is no prior work of MT on the Asian low-resource languages, which open gaps for MT especially on these languages. In this paper, we conducted an empirical study of leveraging additional resources to improve MT on the Asian low-resource languages of translation from Japanese, Indonesian, and Malay to Vietnamese. We focus on two strategies: building bilingual corpora from comparable data and pivot translation, which have been shown to be effectiveness in several languages; however there is no investigation on the Asian low-resource languages. For the first strategy, we built bilingual corpora from Wikipedia parallel titles to enhance training data for MT on the low-resource languages. For the second strategy, we adapted the phrase pivot translation approach to take advantage of existing

bilingual corpora of the languages paired with English, which typically exist due to the popularity of English. Additionally, we introduced a combined model that exploit the two additional resources in a unique system based on a robust linear interpolation using perplexity optimization. Experimental results show that our systems achieve significant improvement on all three language pairs from +2 BLEU up to +7 BLEU points, which confirm the effectiveness of the strategies, our extracted corpus, and the combined model for the task of improving MT on the low-resource languages. Although promising results are achieved, several limitations still remain such as the quality of the extracted corpus or the reordering aspect in Japanese-Vietnamese. For other researchers in the field, this work opens a promising direction for the development of MT on the Asian low-resource languages, which are widely used in the world. There is a big gap between the performance of our results and the powerful Google translation. However, even on the powerful Google translation, the performance of the low-resource languages such as Japanese-Vietnamese is still low. Improving translation quality on such language pairs is needed to address in future research.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 355–362.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 1–44.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of ACL*. Association for Computational Linguistics, 169–176.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2011. Bootstrapping Arabic-Italian SMT through comparable texts and pivot translation. In *Proceedings of EAMT*.

M. Cettolo, C. Girardi, and M. Federico. 2012a. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way (Eds.). 261–268. http://www.mt-archive.info/EAMT-2012-Cettolo

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012b. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*. 261–268.

Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of ACL*. Association for Computational Linguistics, 9–16.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of HLT/NAACL*. Association for Computational Linguistics, 427–436.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a Chinese-Japanese parallel corpus from Wikipedia.. In *Proceedings of LREC*. 642–647.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese–Japanese Wikipedia. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 15, 2, Article 10 (Dec. 2015), 22 pages. DOI:http://dx.doi.org/10.1145/2833089

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: making effective use of multi-parallel Corpora. In *Proceedings of ACL*. Association for Computational Linguistics, 728–735.

Adrià De Gispert and Jose B Marino. 2006. Catalan-English statistical machine translation without parallel corpus: bridging through Spanish. In *Proceedings of LREC*. Citeseer, 65–68.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of ACL*. Association for Computational Linguistics, 412–418.

William A Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19, 1 (1993), 75–102.

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation.. In *ACL (Short Papers)* (2010-06-04). The Association for Computer Linguistics, 57–60.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 187–197.

Ann Irvine. 2013. Statistical machine translation in low resource settings.. In *Proceedings of HLT/NAACL*. Association for Computational Linguistics, 54–61.

MA Jeff, Spyros Matsoukas, and Richard Schwartz. 2011. Improving low-resource statistical machine translation with a novel semantic word clustering algorithm. *Proceedings of the MT Summit XIII* (2011).

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Computational Linguistics* 19, 1 (1993), 121–142.

Sungchul Kim, Kristina Toutanova, and Hwanjo Yu. 2012. Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 694–702.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*. Barcelona, Spain.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*. Phuket, Thailand. http://mt-archive.info/MTS-2005-Koehn.pdf

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for Europe. In *Proceedings of the MT Summit XII*. International Association for Machine Translation.

Philipp Koehn, Barry Haddow, Philip Williams, and Hieu Hoang. 2010. More linguistic annotation for statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, 115–120.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, and others. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*. Association for Computational Linguistics, 177–180.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2013. Improving statistical machine translation by adapting translation models to translationese. *Computational Linguistics* 39, 4 (2013), 999–1023.

Bo Li and Juan Liu. 2008. Mining Chinese-English Parallel Corpora from the Web. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*. http://www.newdesign.aclweb.org/anthology-new/I/I08/I08-2120.pdf

Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation.. In *ACL (1)*. The Association for Computer Linguistics, 11–19.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC*. 489–492.

Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. 2011. BP2EP - Adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of EAMT*. 129–136.

I Dan Melamed. 1996. A geometric approach to mapping bitext correspondence. In *Proceedings EMNLP*. Association for Computational Linguistics.

Robert C Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of AMTA*. 135–144.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 81–88.

Ahmad Musleh, Nadir Durrani, Irina Temnikova, Preslav Nakov, Stephan Vogel, and Osama Alsaad. 2016. Enabling medical translation for low-resource languages. In *Proceedings of CICLing*.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt. (2011).

Malte Nuhn, Arne Mauser, and Hermann Ney. 2012. Deciphering foreign language by combining language models and context vectors. In *Proceedings of ACL*. Association for Computational Linguistics, 156–164.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29, 1 (2003), 19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*. Association for Computational Linguistics, 311–318.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of ACL: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 12–21.

Majid Razmara, Maryam Siahbani, Reza Haffari, and Anoop Sarkar. 2013. Graph Propagation for Paraphrasing Out-of-Vocabulary Words in Statistical Machine Translation.. In *ACL (1)*. The Association for Computer Linguistics, 1105–1115.

Philip Resnik. 1999. Mining the Web for Bilingual Text. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL)*. http://acl.ldc.upenn.edu/P/P99/P99-1068.pdf

Avneesh Saluja, Hany Hassan, Kristina Toutanova, and Chris Quirk. 2014. Graph-based semi-supervised Learning of translation models from monolingual data. In *Proceedings of ACL*. Association for Computational Linguistics, 676–686.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EAMT*. 539–549.

Anil Kumar Singh and Samar Husain. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. In *Proceedings of the ACL Workshop on Building and using Parallel texts*. Association for Computational Linguistics, 99–106.

Dan Ştefănescu and Radu Ion. 2013. Parallel-Wiki: A collection of parallel sentences extracted from Wikipedia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2013)*. 24–30.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the asian language treebank (alt). In *Proceedings of LREC*. 1574–1578.

Hai-Long Trieu, Thanh-Quyen Dang, Phuong-Thai Nguyen, and Le-Minh Nguyen. 2015. The JAIST-UET-MITI machine translation systems for IWSLT 2015. In *The 12th International Workshop on Spoken Language Translation (IWSLT)*.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Erhard Hinrichs and Dan Roth (Eds.). 72–79. http://www.aclweb.org/anthology/P03-1010.pdf

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of HLT/NAACL*. Association for Computational Linguistics, 484–491.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam studies in the theory and history of linguistic science series 4* 292 (2007), 247.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics* (2016).

George Weber. 2008. Top languages. *The World's* 10 (2008).

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings ACL*. Association for Computational Linguistics, 80–87.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of ACL*. Association for Computational Linguistics, 856–863.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR* abs/1609.08144 (2016). http://arxiv.org/abs/1609.08144

Keiji Yasuda, Ruiqiang Zhang, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Method of Selecting Training Data to Build a Compact and Efficient Translation Model.. In *IJCNLP*. 655–660.