A Multilingual Parallel Corpus for Improving Machine Translation on Southeast Asian Languages

Hai-Long Trieu, Le-Minh Nguyen {trieulh, nguyenml}@jaist.ac.jp
Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan

Abstract

Current machine translation systems require large bilingual corpora for training data. With large bilingual corpora, phrase-based and neural-based methods can achieve state-of-the-art performance. Nevertheless, such large bilingual corpora are unavailable for most language pairs called low-resource languages, which causes a bottleneck for the development of machine translation on such languages. For Southeast Asian region, there is a large population with more than five hundred millions people and several languages that can be used popularly in the world, but there are few parallel data for such language pairs. In this work, we built a multilingual parallel corpus for several Southeast Asian languages. Wikipedia articles' titles and inter-language link records were used to extract parallel titles. Parallel articles were collected based on the parallel titles. For each article pair, parallel sentences were extracted based on a length-based and word correspondences sentence alignment method. A multilingual parallel corpus were built with more than 1.1 million parallel sentences of ten language pairs of Indonesian, Malay, Filipino, Vietnamese and the languages paired with English. Experiments were conducted on the Asian Language Treebank corpus and showed the promising performance. Additionally, the corpus was utilized for the IWSLT 2015 machine translation shared task on English-Vietnamese and achieved a significant improvement with +1.7 BLEU point on phrase-based systems and +4.5 BLEU point on a state-of-the-art neural-based system. The corpus can be used to improve machine translation and enhance the development of machine translation on the low-resource Southeast Asian languages.

1 Introduction

Current machine translation (MT) systems require large bilingual corpora for training data. With large bilingual corpora up to millions of parallel sentences, MT systems achieve the state-of-the-art performance on both phrase-based (Bojar et al., 2013) and neural-based (Sennrich et al., 2016a) methods. Such large bilingual corpora are available on several language pairs such as English-German, English-French, Czech-English, Chinese-English. For low-resource language pairs, which are most of languages in the world (Irvine, 2013; Wang et al., 2016), there are only small bilingual corpora available. This causes a bottleneck for MT on such language pairs.

In order to overcome the problem, previous works have made efforts in building bilingual corpora from webs such as in (Utiyama and Isahara, 2003; Li and Liu, 2008; Cettolo et al., 2012). The parallel corpora can be extracted from comparable data such as Wikipedia ((Ştefănescu and Ion, 2013; Chu et al., 2015). The previous work contributed for building bilingual corpora automatically for several low-resource language pairs. For Southeast Asian

languages, there are few bilingual corpora on the languages although there are a high population with more than five hundred millions of people, and there are several languages that can be used popularly in the world such as Indonesian (ranked 12), Vietnamese (ranked 17) as the most popularly used languages (Weber, 2008). This causes an issue for the development of machine translation on the language pairs.

In this work, we built a multilingual parallel corpus to improve machine translation for Southeast Asian languages, which there is no large bilingual corpora. Parallel titles of Wikipedia articles were extracted based on the articles' titles and inter-language link records from the Wikipedia database. Parallel articles were collected based on the parallel titles. Then, parallel sentences were aligned based on a sentence alignment method that is the combination of length-based and word correspondences. A multilingual parallel corpus was built for several low-resource Southeast Asian languages that included more than 1.1 million parallel sentences of ten language pairs between Indonesian, Filipino, Malay, Vietnamese and these languages paired with English. Experiments of machine translation were conducted on the Asian Language Treebank corpus (Thu et al., 2016). Experimental results showed that using the extracted corpus to build machine translation systems can achieve promising results although there is no direct bilingual corpora. Furthermore, experiments were conducted on the IWSLT 2015 machine translation shared task (Cettolo et al., 2015) using the extracted corpus for English-Vietnamese trained on phrase-based and neural-based machine translation systems. Experimental results showed that using the extracted corpus achieved significant improvement in both phrase-based systems and neural-based systems. The corpus can be used to improve machine translation performance and enhance the development of machine translation for the Southeast Asian languages. We released the extracted corpus and the code to build the corpus, which are available at the repository.[1]

We briefly discuss related work in Section 2. The procedures to built the corpus are described in detail in Section 3. The statistics of the extracted corpus are presented in Section 4. In order to effectively utilize the corpus, we present several strategies to exploit the corpus for machine translation in Section 5. Experiments are described in Section 6 to evaluate and utilize the corpus. Conclusions are drawn in Section 7.

2 Related Work

Building parallel corpora from webs 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. (2012), 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 Southeast Asian languages, there are few bilingual corpora. A multilingual parallel corpus was built manually in Thu et al. (2016). The corpus 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 corpora becomes an essential task for the development of natural lan-

[1] https://github.com/nguyenlab/Multi-Wiki

guage processing for the languages including cross-language tasks like machine translation. In our work, a multilingual parallel corpus of several Southeast Asian languages was built. 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 Southeast Asian low-resource languages, in which there has been no work investigated on this task to our best knowledge.

3 Methods

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 crawled based on the parallel titles. Finally, sentences in the article pairs were aligned to extract parallel sentences. We describe these steps in more detail in this section.

3.1 Extracting Parallel Titles

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

No.	Data	page (KB)	langlinks (KB)
1	en	1,477,861	280,617
2	vi	92,541	111,420
3	id	57,921	72,117
4	ms	21,791	56,173
5	fil	5,907	23,446

Table 1: 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 collected the resources for languages: **en** (English), **id** (Indonesian), **fil** (Filipino), **ms** (Malay), and **vi** (Vietnamese); we used the database that was *updated on 2017-01-20*.

We aim to build a multilingual parallel corpus for several low-resource Southeast Asian languages including Indonesian, Malay, Filipino, and Vietnamese, which there are few bilingual corpora. Furthermore, bilingual corpora of the languages paired with English are also important resources for further research including machine translation. Therefore, we collected the Wikipedia database dumps of the five languages: English, Indonesian, Malay, Filipino, and Vietnamese. Table 1 presents the Wikipedia database dumps that we used to extract parallel titles. The English database contains a much larger information in both the articles' titles and the interlanguage link records. Meanwhile, the Filipino database is much smaller, that affects the number of extracted parallel titles as well as final extracted parallel sentences. The extracted parallel titles are presented in Table 2.

[2] https://dumps.wikimedia.org/backup-index.html

No.	Data	Title pairs	Crawled Src Art.	Crawled Trg Art.	Art. Pairs	Src Sent.	Trg Sent.
1	en-id	198,629	197,220	190,954	150,759	4,646,453	990,661
2	en-fil	52,749	51,698	51,157	50,021	3,428,599	367,276
3	en-ms	204,833	201,688	199,950	160,709	2,158,726	320,624
4	en-vi	452,415	433,124	436,488	420,919	12,130,133	3,831,948
5	id-fil	30,313	29,961	24,946	22,760	502,457	254,216
6	id-ms	98,305	88,028	89,936	68,676	452,604	403,807
7	id-vi	159,247	149,974	128,530	121,673	1,201,848	1,878,855
8	fil-ms	25,231	21,856	25,023	21,135	202,851	243,361
9	fil-vi	36,186	30,540	35,625	28,830	267,453	723,155
10	ms-vi	133,651	118,647	116,620	105,692	560,042	1,256,468

Table 2: Extracted and processed data from parallel titles; **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).

3.2 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 from punctuations. The sentence and word tokenization steps were conducted using the Moses scripts.[3]

As described in Table 2, using the title pairs, we obtained a high ratio of crawled articles. For instance, using 198k title pairs of English-Indonesian, we crawled 197k English articles and 190k Indonesian articles successfully, which there existed the article based on a title. This issue is emphasized because sometimes there is no existed article given a title that will show an error in crawling. For the case of Indonesian-Vietnamese, although there was 159k extracted parallel titles, we obtained 128k Vietnamese articles, which there were more than 30k error or inexistent articles given the set of titles.

3.3 Aligning Parallel Sentences

Sentence alignment is an essential task in building parallel corpora. In 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 combination of length-based and word-based (Moore, 2002; Varga et al., 2007), the hybrid method of Moore (2002) achieved the best performance compared with other sentence alignment approaches as the evaluation of Singh and Husain (2005). In our work, for each parallel article pair, we aligned sentences using the Microsoft bilingual sentence aligner (Moore, 2002). There are several reasons to adapt 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 languages in the Southeast Asian including Indonesian, Malay,

[3] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer

Vietnamese, Filipino, and the languages paired with English are not closed languages exception for the Indonesian-Malay. Second, since the Wikipedia bilingual articles are the kind of comparable data, it varies 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.

Let l_s and l_t be the lengths of source and target sentences, respectively. Then, l_s and l_t varies according to Poisson distribution as follows:

$$P(l_t|l_s) = exp^{-l_t r}\frac{(l_s r)^{l_t}}{l_t!} \quad (1)$$

Where *r* is the ratio of the mean length of target sentences to the mean length of source sentences. As shown in the method of Moore (2002), the length-based phase based on the Poisson distribution

Sentence pairs extracted from the length-based phase are then used to train IBM Model 1 (Brown et al., 1993) to build a bilingual dictionary. The dictionary was then combined with the length-based phase to produce final alignments, which are described as follows:

$$P(s,t) = \frac{P_{1-1}(l_s,l_t)}{(l_s+1)^{l_t}}(\prod_{j=1}^{l_t}\sum_{i=0}^{l_s} tr(t_j|s_i))(\sum_{i=1}^{l_e} f_u(e_i)) \quad (2)$$

Where: $tr(t_j|s_i)$ is the probability of the word pair $(t_j|s_i)$ trained by IBM Model 1; f_u is the observed relative unigram frequency of the word in the text in the corresponding language.

Challenges in aligning Wikipedia articles As we discussed above, the Wikipedia article pairs greatly vary in terms of sentence length in the article pairs because of this kind of comparable data. Furthermore, in some article pairs, the articles in two languages even contain many differences in content, priorities, interests, and bias of the authors, groups or countries involved, etc. Such differences cause many challenges for aligning Wikipedia articles to create a parallel corpus. For our first effort in building this corpus, we used the hybrid sentence alignment method to extract sentence pairs for the first version of this corpus without any strategy to filter or extract parallel sentences in dealing with these challenges. We plan to conduct further analysis as well as strategies to deal with the challenges and improve the quality of this corpus in future work. A method proposed in Munteanu and Marcu (2006) can be utilized for this task, in which parallel sub-sentential fragments are extracted from comparable data.

4 Extracted Corpus

We obtained a multilingual parallel corpus of ten language pairs, which are among Southeast Asian languages and the languages paired with English as described in Table 3. In totally, the corpus contains a huge number of parallel sentences up to more than 1.1 million sentence pairs which can be valuable when there is no available bilingual corpora for almost such language pairs. Large bilingual corpora can be extracted such as: English-Vietnamese (408k parallel sentences), Indonesian-English (234k parallel sentences). However, because of the smaller number of the input parallel articles for several language pairs, a much smaller number of parallel sentences were extracted like Indonesian-Filipino (9k) and Filipino-English (22k).

Furthermore, we extracted monolingual data sets for the languages: Indonesian, Malay, Filipino, and Vietnamese, which are almost publicly unavailable. The data sets are described in Table 4. Large monolingual data sets were obtained such as Indonesian (3.1 million sentences), Malay (1.5 million sentences), and Vietnamese (up to 7.6 million sentences). The data sets are useful for such low-resource languages such as training language models and other tasks.

No.	Data	Sent. Pairs	Src Words	Trg Words	Src Vocab.	Trg Vocab.
1	en-id	234,380	4,648,359	4,359,976	208,920	209,859
2	en-fil	22,758	447,719	399,058	42,670	44,809
3	en-ms	198,087	3,273,943	3,221,738	156,806	148,133
4	en-vi	408,552	7,229,963	8,373,549	274,178	222,068
5	id-fil	9,952	132,097	172,363	18,531	19,737
6	id-ms	83,557	1,464,506	1,447,247	87,240	92,126
7	id-vi	76,863	1,014,351	1,136,710	67,211	57,788
8	fil-ms	4,919	78,729	66,324	10,184	10,671
9	fil-vi	10,418	141,135	151,086	15,641	13,071
10	ms-vi	65,177	928,205	896,784	60,574	52,673
	Total	1,114,663	–	–	–	–

Table 3: Extracted Southeast Asian multilingual parallel corpus

Data set	Sentences	Vocab.	Size (KB)
id	3,147,570	917,861	369
fil	1,034,215	252,565	113
ms	1,527,834	599,396	172
vi	7,690,426	936,137	1,033

Table 4: Monolingual data sets

5 Domain Adaptation

The question now is that how can we utilize the corpus effectively. If there are existing bilingual corpora for the language pairs, which strategies we can use to combine and take advantage the full potential of the corpus. We discuss the issue of domain adaptation about the strategies to combine bilingual corpora in this section.

We assume that given a language pair, there exist a bilingual corpus called the *direct corpus*. The corpus extracted from Wikipedia can be used as an additional resource, called the *alignment corpus*. For phrase-based machine translation (Koehn et al., 2003), a bilingual corpus is used to train a phrase table. We used the *direct corpus* and the *alignment corpus* to generate two phrase tables called the *direct* and the *alignment* components. The two components were combined using the linear interpolation as described in Equation 3.

$$p(t|s) = \lambda_d p_d(t|s) + \lambda_a p_a(t|s) \tag{3}$$

where $p_d(t|s)$ and $p_a(t|s)$ stand for the translation probabilities of the *direct* and the *alignment* models, respectively; interpolation parameters: λ_d and λ_a (where $\lambda_d + \lambda_a$=1).

We adapted the linear interpolation (Sennrich, 2012), which is a robust method for a weighted combination of translation models. Specifically, we used two strategies called *tune* and *weights*.

- *tune*: a tuning set was used; λ_d and λ_a were calculated as the weights that minimize cross-entropy on the tuning set using the setting *combine_given_tuning_set* (Sennrich, 2012). [4]
- *weights*: The two translation models were first used for decoding the tuning set separately to generate two BLEU scores. Then, the interpolation weights were set using the ratios of the two BLEU scores using the setting *combine_given_weights* Sennrich (2012).

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

6 Experiments on Machine Translation

The parallel corpus extracted from Wikipedia was then used for training SMT models. We aim to exploit the data to improve SMT on low-resource languages.

6.1 SMT on the Asian Language Treebank Parallel Corpus

6.1.1 Training Data

We evaluate the corpus on SMT experiments. For development and testing data, we used the ALT corpus (Asian Language Treebank Parallel Corpus) Thu et al. (2016), this is a corpus including 20K multilingual sentences of English, Japanese, Indonesian, Filipino, Malay, Vietnamese, and some other Southeast Asia languages. We extracted the development and test sets from the ALT corpus: 2k sentence pairs for development sets, and 2k sentence pairs for test sets.

6.1.2 Training Details

We trained SMT models on the parallel corpus using the Moses toolkit (Koehn et al., 2007). The word alignment was trained using GIZA++ (Och and Ney, 2003) with the configuration *grow-diag-final-and*. A 5-gram language model of the target language was trained using KenLM (Heafield, 2011). For tuning, we used batch MIRA (Cherry and Foster, 2012). For evaluation, we used the BLEU scores (Papineni et al., 2002) based on the *multi-bleu.perl* script; the development sets, test sets, and scripts to calculate the BLEU scores are also available in the repository of this paper.

6.1.3 Results

Table 5 describes the experimental results on the development and test sets. It is noticeable that the SMT models trained on the bilingual data aligned from Wikipedia produced promising results.

No.	Language Pairs	Dev (L1-L2)	Test (L1-L2)	Dev (L2-L1)	Test (L2-L1)
1	en-id	30.56	28.87	30.14	29.01
2	en-fil	18.54	19.08	18.98	19.89
3	en-ms	29.85	33.23	28.87	23.82
4	en-vi	30.58	34.42	23.01	22.56
5	id-fil	11.36	11.04	9.58	9.77
6	id-ms	31.64	30.21	31.56	30.11
7	id-vi	21.85	22.42	17.41	17.45
8	fil-ms	7.43	8.02	8.70	9.27
9	fil-vi	5.97	6.69	6.45	7.15
10	ms-vi	15.51	18.12	11.96	13.88

Table 5: Experimental results on the development and test sets (BLEU); **Dev (L1-L2)**, **Test (L1-L2)**, **fil-ms**: the translation scores on the development (test) set of the translation from the first language (**L1(fil)**) to the second language (**L2 (ms)**) in the language pair **fil-ms**; **Dev (L2-L1)**, **Test (L2-L1)**, **fil-ms**: the translation on the development (test) set of the inverse translation (from **ms** to **fil**)

For the results on the development sets, we achieved promising results with high BLEU points such as: the Indonesian-Malay pairs (Indonesian-Malay 31.64 BLEU points, Malay-Indonesian 31.56 BLEU points). Similarly, several language pairs also showed high BLEU points such as: English-Vietnamese (30.58 and 23.01 BLEU points), English-Malay (29.85

and 28.87 BLEU points), English-Indonesian (30.56 and 30.14 BLEU points), and Indonesian-Vietnamese (21.85 and 17.41 BLEU points). The language pairs which showed high scores contain a large number of sentences, for instance English-Vietnamese (408k sentence pairs), English-Indonesian (234k sentence pairs), and English-Malay (198k sentence pairs). Nevertheless, since the small number of the extracted corpus on several languages paired with Filipino such as Indonesian-Filipino (9.9k sentence pairs), Malay-Filipino (21.1k sentence pairs), and Vietnamese-Filipino (10.4k sentence pairs), the experimental results showed much lower performance than other language pairs: Indonesian-Filipino (11.36 and 9.58 BLEU points), Malay-Filipino (8.70 and 7.43 BLEU points), and Vietnamese-Filipino (6.45 and 5.97 BLEU points).

Similarly, for the experimental results on the test sets, the language pairs with large bilingual corpora achieved high performance: English-Indonesian (28.87 and 29.01 BLEU points), English-Malay (33.23 and 23.82 BLEU points), English-Vietnamese (34.42 and 22.56 BLEU points). The situation of languages paired Filipino showed the much lower performance: Indonesian-Filipino (11.04 and 9.77 BLEU points), Malay-Filipino (9.27 and 8.02 BLEU points), and Vietnamese-Filipino (7.15 and 6.69 BLEU points).

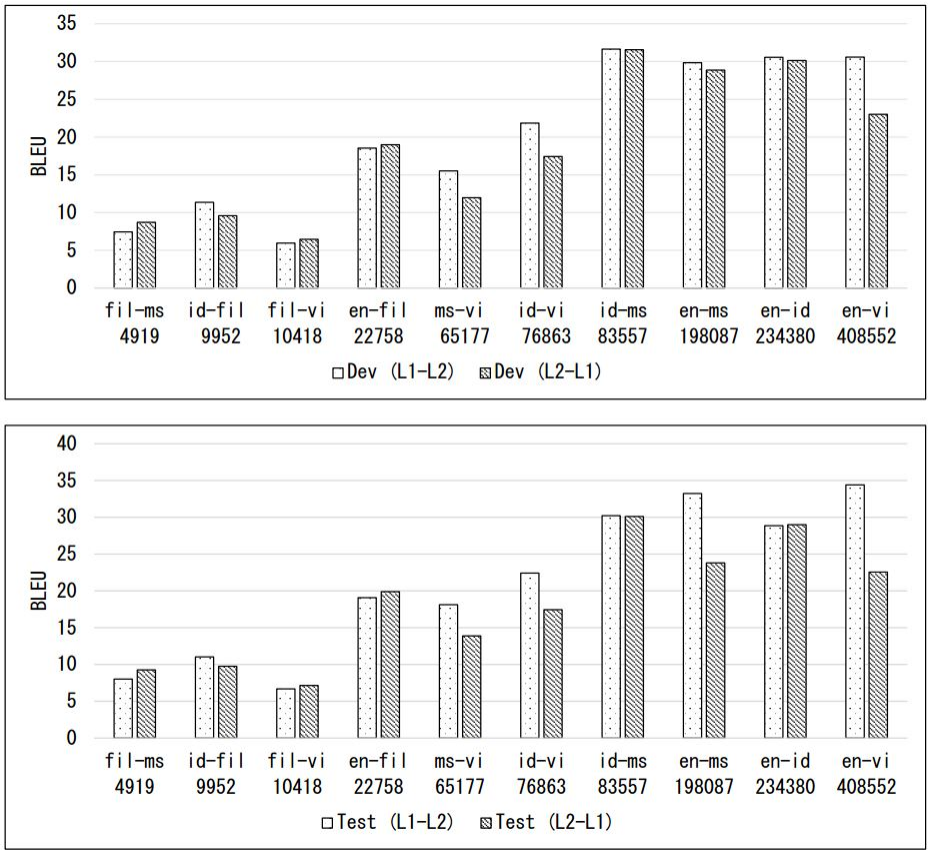

Figure 1: Experimental results on the development and test sets; the corpus's size is presented for each language pair (**fil-ms 4919**: the Filipino-Malay corpus with 4,919 parallel sentences)

Figure 1 presents experimental results on the development sets (test sets) that vary in

several aspects: the translation directions (L1-L2, L2-L1), the corpus's size, and the language pairs. There are several interesting findings from the charts. First, the bigger the corpus's size, the higher the BLEU scores. We sorted the corpus's size increasingly from the left to right. For instance, since the corpora' sizes of language pairs such as Filipino-Malay (4.9k), Indonesian-Filipino (9.9k), and Filipino-Vietnamese (10.4k) are much smaller than that of the language pairs such as Indonesian-Malay (83.5k), English-Indonesian (234k), English-Vietnamese (408k), the BLEU scores also show the correlation of the two language-pair groups: Filipino-Malay, Indonesian-Filipino, Filipino-Vietnamese ($<$10 or $\approx$ 10 BLEU points); Indonesian-Malay, English-Indonesian, English-Vietnamese ($\approx$ 25-30 BLEU points). Second, in the aspect of the translation directions (L1-L2, L2-L1), the scores of the two translations in each language pair are mostly similar to each other in most cases, for instance: English-Indonesian (30.56 and 30.14 BLEU points in the two translation directions on the development set, 28.87 and 29.01 on the test set), Indonesian-Malay (31.64 and 31.56 BLEU points on the development set, 30.21 and 30.11 on the test set). Nevertheless, for Vietnamese, the translation scores from a language to Vietnamese are much higher than the translation scores from Vietnamese to that language in most cases, for instance: Malay-Vietnamese (15.51 BLEU point (ms-vi) vs. 11.96 (vi-ms) on the development set, 18.12 (ms-vi) vs. 13.88 (vi-ms) on the test set), Indonesian-Vietnamese (21.85 vs. 17.41 BLEU points on the development set, 22.42 vs. 17.45 BLEU points on the test set), and English-Vietnamese (30.58 vs. 23.01 BLEU points on the development set, 34.42 vs. 22.56 BLEU points on the test set). This problem of the unbalance scores between the two translation directions of a language paired with Vietnamese as well as other language pairs should be further investigated.

6.2 Evaluation on the IWSLT 2015 Machine Translation Shared Task

In this section, we evaluated the extracted corpus on the IWSLT 2015 machine translation shared task on English-Vietnamese. We aim to evaluate whether the *Wikipedia* corpus can improve some baseline systems on the shared task. In addition, we conducted various experiments of the domain adaptation strategies, statistical machine translation, and neural machine translation using the *Wikipedia* corpus to explore optimal strategies in exploiting the corpus.

6.2.1 Training Data

Data	Sentences	Src Words	Trg Words	Src Vocab.	Trg Vocab.
constrained	131,019	2,534,498	2,373,965	50,118	54,565
unconstrained	456,350	8,485,112	8,132,913	114,161	124,846
constrained+Wikipedia	538,981	9,710,389	9,017,601	288,785	345,839
unconstrained+Wikipedia	864,312	15,661,003	14,776,549	338,424	403,581
tst2012	1,581	28,773	27,101	3,713	3,958
tst2013	1,304	28,036	27,264	3,918	4,316
tst2015	1,080	20,844	19,951	3,175	3,528

Table 6: Data sets on the IWSLT 2015 experiments; **Src Words (Trg Words)**: the number of words in the source (target) side of the corpus; **Src Vocab. (Trg Vocab.)**: the vocabulary size in the source (target) side of the corpus

We used the data sets provided by the International Workshop on Spoken Language Translation (IWSLT 2015) machine translation shared task (Cettolo et al., 2015), which include three data sets of the training, development, and test sets extracted from subtitles of TED talks.[5] For

[5]https://www.ted.com/talks

the training data, the data set called the *constrained* data of 131k parallel sentences. The workshop provided data sets for development and test sets: *tst2012*, *tst2013*, and *tst2015*. In all experiments, we used the *tst2012* for the development set, the *tst2013* and *tst2015* for the test sets.

In addition, we used two other data sets for training data: the corpus of National project VLSP (Vietnamese Language and Speech Processing)[6] and the EVBCorpus (Ngo et al., 2013). The two data sets were merged with the *constrained* data to obtain a large training data set called the *unconstrained* data. The training, development, and test sets are described in Table 6.

6.2.2 Training Details

We trained translation systems using two methods: SMT and NMT.

Statistical Machine Translation In order to train SMT models, we used the well-known Moses toolkit (Koehn et al., 2007). The GIZA++ (Och and Ney, 2003) was used to train word alignment. For language model, we used KenLM (Heafield, 2011) to train 5-gram language models on the target side (Vietnamese) of the training data sets. The parameters were tuned using batch MIRA (Cherry and Foster, 2012). BLEU (Papineni et al., 2002) was used as the metric for evaluation.

Neural Machine Translation In our work, we based on the model of Sennrich et al. (2016a), which are encoder-decoder networks with an attention mechanism (Bahdanau et al., 2015). For NMT model, we adopted the attentional encoder-decoder networks combined with byte-pair encoding (Sennrich et al., 2016a). In our experiments, we set the word embedding size 500, and hidden layers size of 1024. Sentences are filtered with the maximum length of 50 words. The minibatches size is set to 60. The models were trained with the optimizer Adadelta (Zeiler, 2012). The models were validated each 3000 minibatches based on the BLEU scores on development sets. We saved the models for each 6000 minibatches. For decoding, we used beam search with the beam size of 12. We trained NMT models on an Nvidia GRID K520 GPU.

6.2.3 Results

Model	**tst2012**	**tst2013**	**tst2015**
Wikipedia	18.40	22.06	20.34
constrained	24.72	27.31	25.47
constrained+Wikipedia	24.78	27.89	26.69
constrained*Wikipedia (tune)	24.65	28.05	27.00
constrained*Wikipedia (weights)	**24.95 (+0.23)**	**28.51 (+1.20)**	**27.21 (+1.74)**
unconstrained	34.42	27.19	25.41
unconstrained+Wikipedia	33.88	27.28	26.36
unconstrained*Wikipedia (tune)	34.44	27.55	26.68
unconstrained*Wikipedia (weights)	**34.73 (+0.31)**	**28.04 (+0.85)**	**26.78 (+1.37)**

Table 7: Experimental results using phrase-based statistical machine translation; *constrained+Wikipedia*: the *constrained* data was merged with the *Wikipedia* corpus; *unconstrained*Wikipedia*: interpolation of the two models; *tune, weights*: the two interpolation settings; the **bold** indicates the best results for each setup

SMT results Table 7 presents experimental results using SMT models. Using the *Wikipedia* corpus, we achieved promising results: 18.40 BLEU point (*tst2012*), 22.06 (*tst2013*), and 20.34 (*tst2015*). When the *Wikipedia* corpus was merged with the *constrained* data for training data,

[6] http://vlsp.vietlp.org:8080/demo/?page=home

a significant improvement was achieved especially on the *tst2015* (26.69 BLEU point, which improved 1.22 BLEU point from the model using the *constrained* data). Nevertheless, the domain adaptation strategies show even better performance than the merging setting, in which the *weights* setting model obtained the best performance with +1.74 BLEU point improvement on the *tst2015*.

NMT results The NMT results are described in Table 8. From the experimental results, we can observe that the systems obtain the higher scores when the size of training data sets increase (from the *Wikipedia*, *constrained*, *unconstrained* to the merging in which the *unconstrained* data was merged with the *Wikipedia* corpus). It is interesting to note that using the *Wikipedia* corpus to enhance the translation systems trained on existed data sets based on NMT achieved the significant improvement up to +4.51 BLEU points on the *tst2015*.

Model	**tst2012**	**tst2013**	**tst2015**
constrained	20.21	23.59	17.27
Wikipedia	15.29	18.43	17.58
unconstrained	24.05	26.71	22.30
unconstrained+Wikipedia	**25.29 (+1.24)**	**28.93 (+2.21)**	**26.81 (+4.51)**

Table 8: Experimental results on neural machine translation (NMT) ; the **bold** indicates the best results for each setup

A work that enhanced neural machine translation using additional data is presented in Sennrich et al. (2016b) called back-translation. In the back-translation method, a synthetic corpus is generated by translating a large monolingual data in a target language into source sentences. For further evaluation and utilization of the extracted *Wikipedia* corpus, a comparison and adaptation the back-translation method is needed in future work.

SMT vs. NMT We compared the improvement of the *Wikipedia* corpus using the SMT versus NMT systems. Experimental results showed that the SMT systems obtained better performance on the *unconstrained* data (456k): 25.41 vs. 22.30 on the *tst2015*. Nevertheless, when the *Wikipedia* corpus was utilized, which was merged with the *unconstrained* data to enlarge the training data (864k), the NMT systems outperformed the SMT systems, which indicates the benefit when utilizing the *Wikipedia* corpus on NMT compared with SMT systems. Table 9 presents the comparison in more detail.

Model	**tst2012**	**tst2013**	**tst2015**
SMT systems			
unconstrained	34.42	27.19	25.41
unconstrained+Wikipedia	33.88	27.28 (+0.09)	26.36 (+0.95)
unconstrained*Wikipedia (tune)	34.44 (+0.02)	27.55 (+0.36)	26.68 (+1.27)
unconstrained*Wikipedia (weights)	34.73 (+0.31)	28.04 (+0.85)	26.78 (+1.37)
NMT systems			
unconstrained	24.05	26.71	22.30
unconstrained+Wikipedia	25.29 (+1.24)	28.93 (+2.21)	26.81 (+4.51)

Table 9: SMT versus NMT in using the *Wikipedia* corpus

From this comparison, we investigated the strategies to utilize the *Wikipedia* corpus most effectively for improving machine translation on low-resource languages, in which the corpus was utilized more effectively when using the NMT models.

7 Conclusion

Current machine translation systems in both phrase-based and neural-based methods require large bilingual corpora for training data. Nevertheless, such large bilingual corpora are unavailable for most language pairs called low-resource languages. This causes a bottleneck for the languages. In Southeast Asian languages, although there are a high population with more than five hundred millions of people, and there are several languages that can be used popularly in the world like Indonesian, Malay, and Vietnamese, there are few bilingual corpora on these language pairs, which causes a bottleneck for machine translation. In this paper, we introduce building a multilingual parallel corpus for several Southeast Asian languages of Indonesian, Malay, Filipino, Vietnamese, and the languages paired with English to improve machine translation. The corpus was built based on the Wikipedia's parallel titles of articles extracted by the articles' titles and inter-language link records. The parallel titles were used to collect parallel articles. For each article pair, parallel sentences were extracted based on a length-based and word correspondence sentence alignment method. A huge multilingual parallel corpus were obtained with more than 1.1 million parallel sentences of ten language pairs of the Southeast Asian languages. Experiments were conducted on the Asian Language Treebank and showed the promising results. Additionally, the corpus was utilized for the IWSLT 2015 machine translation shared task. A significant improvement was achieved on both phrase-based and neural-based systems with +1.7 and 4.5 BLEU points. The corpus can improve machine translation for the low-resource Southeast Asian languages and contribute to the development of machine translation on the low-resource languages.

Acknowledgement

This work was supported by JSPS KAKENHI Grant number JP15K16048 and the VNU project "Exploiting Very Large Monolingual Corpora for Statistical Machine Translation" (code QG.12.49).

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44. Association for Computational Linguistics.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of ACL*, pages 169–176. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In Cettolo, M., Federico, M., Specia, L., and Way, A., editors, *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M. (2015). The iwslt 2015 evaluation campaign. *Proc. of IWSLT, Da Nang, Vietnam*.

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

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proc. of HLT/NAACL*, pages 427–436. Association for Computational Linguistics.

Chu, C., Nakazawa, T., and Kurohashi, S. (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):10:1–10:22.

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

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

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

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

Kim, S., Toutanova, K., and Yu, H. (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*, pages 694–702. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*, pages 177–180. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics.

Li, B. and Liu, J. (2008). Mining Chinese-English parallel corpora from the web. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *Proceedings of LREC*, pages 489–492.

Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. In *Proceedings EMNLP*. Association for Computational Linguistics.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144. Springer.

Munteanu, D. S. and Marcu, D. (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*, pages 81–88. Association for Computational Linguistics.

Ngo, Q. H., Winiwarter, W., and Wloka, B. (2013). Evbcorpus-a multi-layer english-vietnamese bilingual corpus for studying tasks in comparative linguistics. In *Proceedings of the 11th Workshop on Asian Language Resources (11th ALR within the IJCNLP2013)*, pages 1–9.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318. Association for Computational Linguistics.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL)*.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EAMT*, pages 539–549.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation (WMT)*.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Singh, A. K. and Husain, S. (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*, pages 99–106. Association for Computational Linguistics.

Ştefănescu, D. and Ion, R. (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)*, pages 24–30.

Thu, Y. K., Pa, W. P., Utiyama, M., Finch, A., and Sumita, E. (2016). Introducing the asian language treebank (alt). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1574–1578.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In Hinrichs, E. and Roth, D., editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. *Amsterdam studies in the theory and history of linguistic science series 4*, 292:247.

Wang, P., Nakov, P., and Ng, H. T. (2016). Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics*.

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

Wu, D. (1994). Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 80–87. Association for Computational Linguistics.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *CoRR*.