Noname manuscript No.
(will be inserted by the editor)

MT-WIKI: A Wikipedia-based Multilingual Parallel Corpus for Machine Translation on Low-Resource Languages

Hai-Long Trieu · Ashwin Ittoo

Received: date / Accepted: date

Abstract Large parallel corpora play an essential role in machine translation and other cross-lingual natural language processing tasks. Training high quality translation systems requires large parallel corpora up to million sentence pairs, which are mostly available on European languages and several other languages paired with English. For Southeast Asian languages, there are few or even no existing parallel corpus, which causes the research of machine translation on these languages more challenging. Although there are several efforts in building parallel corpora manually or collecting from available parallel resources, the corpus size is still limited. In this work, we introduce a multilingual parallel corpus containing 2.5 millions parallel sentences on ten language pairs of several Southeast Asian languages among Filipino, Malay, Indonesian, Vietnamese and these languages paired with English. The corpus is automatically built from the abundantly available Wikipedia resource. On the document level alignment, we utilize the available Wikidump inter-language link record information to extract pairs of article titles then article texts. On the sentence level alignment, we employ a powerful and independent sentence aligner to extract parallel sentences. To evaluate the corpus on machine translation, we conducted experiments using the state-of-the-art Transformer model for building translation systems and evaluated on two manually created corpora, which are the Asian Language Treebank and the IWSLT 2015 shared task. Experimental results on various settings show that the corpus can help to significantly improve the translation performance on these low-resource languages. Additionally, the framework to build the corpus is language-independent, which can be extended to build parallel corpus for other language pairs. The corpus and code are publicly available at `https://github.com/trieuhl/mt-wiki`.

Keywords Parallel corpora · machine translation · Southeast Asian languages · Wikipedia

Hai-Long Trieu
Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: trieulh@jaist.ac.jp

Ashwin Ittoo
University of Liège, Belgium
E-mail: ashwin.ittoo@uliege.be

1 Introduction

Large, parallel corpora of translated sentence pairs are a critical resource at the core of machine translation (MT) systems (Koehn et al, 2003; Cho et al, 2014; Vaswani et al, 2017). These corpora exist predominantly for European languages, such as English-German, English-French and Czech-English (Koehn, 2005; Steinberger et al, 2006; Bojar et al, 2016). Research has focused on creating such corpora for Chinese (English-Chinese) (Tian et al, 2014) and Japanese (English-Japanese) (Utiyama and Isahara, 2007) as well. The ready availability of these large parallel corpora have fuelled the recent breakthroughs in MT and have contributed to the state-of-the-art performance of the MT methods for the aforementioned language pairs (Edunov et al, 2018; Ng et al, 2019; Wang et al, 2019).

However, at the same time, MT research on other languages, and in particular, focusing on South East Asian (SEA) languages, such as Filipino, Indonesian, Malay, Vietnamese, has been largely overlooked, and is still lagging behind. One of the major impediments, hindering the development of novel MT methods for these languages, e.g. between Malay-Vietnamese, is the absence of reasonably-large parallel corpora covering these language pairs. Few parallel corpora exist with small size (Tan and Bond, 2011; Riza et al, 2016). In the absence of such large corpora and the lack of substantial training data, MT performance degrades drastically when applied to SEA languages (Singh et al, 2016; Tan, 2016; Wang et al, 2016). Thus, an important research question in the MT community is how to improve the performance of MT methods for low-resource SEA languages? Furthermore, many (low-resource) SEA languages are prominently used by large populations worldwide. For e.g., Indonesian and Vietnamese are the 12th and 17th most widespread languages worldwide as presented by Weber (2008). According to the Ethnologue (2019, 22nd edition),[1] the languages like Indonesian and Vietnamese are included in the top of twenty languages which are used in the world. Thus, besides the scientific/academic aspects, research in improving MT performance on these languages is timely and relevant, with important societal ramifications.

In this article, we make a first, significant step towards in this direction. As our main contribution, we develop and present a large multilingual parallel corpus, with more than 2.5 million sentence pairs, encompassing four SEA languages, viz. Indonesian, Malay, Filipino and Vietnamese. These languages are also paired with English. The corpus is created based on the abundantly available resource Wikidump. From the article title and inter-language link records information of available in the Wikidump, we first extract parallel article titles for each language pair. Article texts are then extracted given the parallel titles. Finally, sentences in each article pair are aligned using an existing language-independent and powerful sentence aligner to create the corpus.

We evaluate the corpus in machine translation task using two high quality manually constructed corpora namely the Asian Language Treebank (ALT) corpus (Riza et al, 2016) and the IWSLT 2015 shared tasks (Cettolo et al, 2015). We employed the powerful Transformer model (Ott et al, 2019) for building MT systems. We showed that an MT system trained solely on our (automatically created) parallel corpus achieved reasonably good performance in translating across the four SEA languages. Our experiments also revealed that the incorporation of our proposed corpus (during training) significantly improved the performance MT systems that had been trained on high-quality, manually curated (but smaller) parallel corpora. The quality of our proposed parallel corpus is comparable to that of (smaller)

[1] `https://www.ethnologue.com`

manually curated corpora, despite the fact that ours has been created automatically from a potentially noisy source.

The contributions of this paper are as follows:

1. We introduce a new corpus for the SEA languages, which are rare and can help to improve MT as well as other multilingual NLP tasks on these languages.
2. We present a simple, but yet very promising, methodology for automatic corpus creation that attempts to leverage on existing resources, such as Wikipedia. The method can be applied to build parallel corpus for any language pair given that the Wikipedia data is available
3. We released the corpus for research in MT as well as in other bilingual tasks in the low-resource SEA languages. Furthermore, we provide the code to reproduce the results in this paper as well as apply to build parallel corpus for any other language pair from the Wikipedia data. The dataset and code are available at `https://github.com/trieuhl/mt-wiki`

2 Related Work

Since large parallel corpora play an essential role in cross-lingual natural language processing tasks such as machine translation, several multilingual corpora have been collected and built including the *Europarl* (Koehn, 2005),[2] *JRC-Acquis* (Steinberger et al, 2006),[3] *UN Parallel Corpus* (Ziemski et al, 2016),[4] *WIT3* corpus (Cettolo et al, 2012),[5] and *OPUS* (Tiedemann, 2012).[6] These corpora are mostly in European languages and collected from various available resources of multilingual texts such as legislative text, parliament proceedings or documents, and video subtitles. There are several other large bilingual corpora in Asia languages such as the *UM-Corpus* English-Chinese (Tian et al, 2014)[7] collected from bilingual websites, and the *NTCIR PatentMT* corpus[8] of Japanese-English and Chinese-English collected from patent description.

Besides building corpora from available parallel resources (Koehn, 2005; Steinberger et al, 2006; Cettolo et al, 2012), automatically extracting parallel corpora from webs is also a direction to build and enlarge parallel corpora (Resnik, 1999). For this direction, building parallel corpora from Wikipedia is also deployed (Adafre and De Rijke, 2006; Smith et al, 2010; Gupta et al, 2013) due to the large amount of Wikipedia articles publicly available on many languages. Many methods have been proposed to extract parallel texts in sentences or phrases from Wikpedia texts such as based on sentence similarity (Adafre and De Rijke, 2006), linked-based method (Mohammadi and GhasemAghaee, 2010), binary and cosine similarity (Saad et al, 2013), cross-lingual information retrieval (Ştefănescu and Ion, 2013), extracting document level alignment with maximum entropy classifier (Smith et al, 2010), clause level alignment (Plamadă and Volk, 2013), extracting parallel fragments (Gupta et al, 2013; Chu et al, 2015). Most of the work is applied on European languages such as English, Dutch, Spanish, Portuguese, German, Polish, Bulgarian (Adafre and De Rijke, 2006;

[2] `https://www.statmt.org/europarl/`

[3] `https://ec.europa.eu/jrc/en/language-technologies`

[4] `https://conferences.unite.un.org/uncorpus/`

[5] `https://wit3.fbk.eu`

[6] `http://opus.nlpl.eu`

[7] `http://nlp2ct.cis.umac.mo/um-corpus/`

[8] `http://ntcir.nii.ac.jp/PatentMT/`

Otero and Lopez, 2010; Barrón-Cedeño et al, 2015; Ştefănescu and Ion, 2013). Several corpora are built on Asia languages paired with English such as Persian-English (Otero and Lopez, 2010), English-Bengali (Gupta et al, 2013), or Chinese-Japanese (Chu et al, 2015). In our framework, we utilized the available Wikipedia dumps' inter-language link information for document-level alignment before extracting parallel sentences based on a language-independent sentence aligner. In terms of languages, our work is the first effort to build a multilingual parallel corpus for the low-resource Southeast Asian languages among Filipino, Indonesian, Malay, Vietnamese, and paired with English.

For Southeast Asian languages, there are several efforts in building parallel corpora (Tan and Bond, 2011; Ngo et al, 2013; Riza et al, 2016). For English-Vietnamese, the *EVBCorpus* (Ngo et al, 2013) contains 800k parallel sentences and its updated version[9] containing 2.2M parallel sentences are collected from bilingual resources such as books, news, and legal texts. For Indonesian-English, the *BPPT* corpus contains more than 300k parallel sentences collected from internet such as national newspapers/magazines and governmental institutions and corrected by professional translators. There are two existing multilingual corpora including the *NTU-MC* corpus (Tan and Bond, 2011) containing 15k sentences in six languages English, Chinese, Japanese, Korean, Indonesian, and Vietnamese and the *Asian Language Treebank (ALT)* corpus (Riza et al, 2016) containing 20k multilingual sentences on English, Filipino, Indonesian, Japanese, Khmer, Laotian, Malay, Myanmar, Thai, and Vietnamese. While the *NTU-MC* corpus is collected from a website of Singapore Tourism Board with parallel texts, the *ALT* corpus is built by manually translating English texts into the other languages. Since most of these work are based on available parallel resources such as bilingual webs or documents or manually translated texts, the corpus size is still limited. For our work, we present a simple framework but can be able to automatically extract a large multilingual parallel corpus on the low-resource Southeast Asian languages, and release the corpus containing more than 2.5M parallel sentences, which can significantly support and improve the research of cross-lingual tasks on these limited available parallel resources.

3 Method

Our framework to build the parallel corpus includes three steps: parallel title extraction, article collection, and sentence alignment. The framework overview is presented in Figure 1.

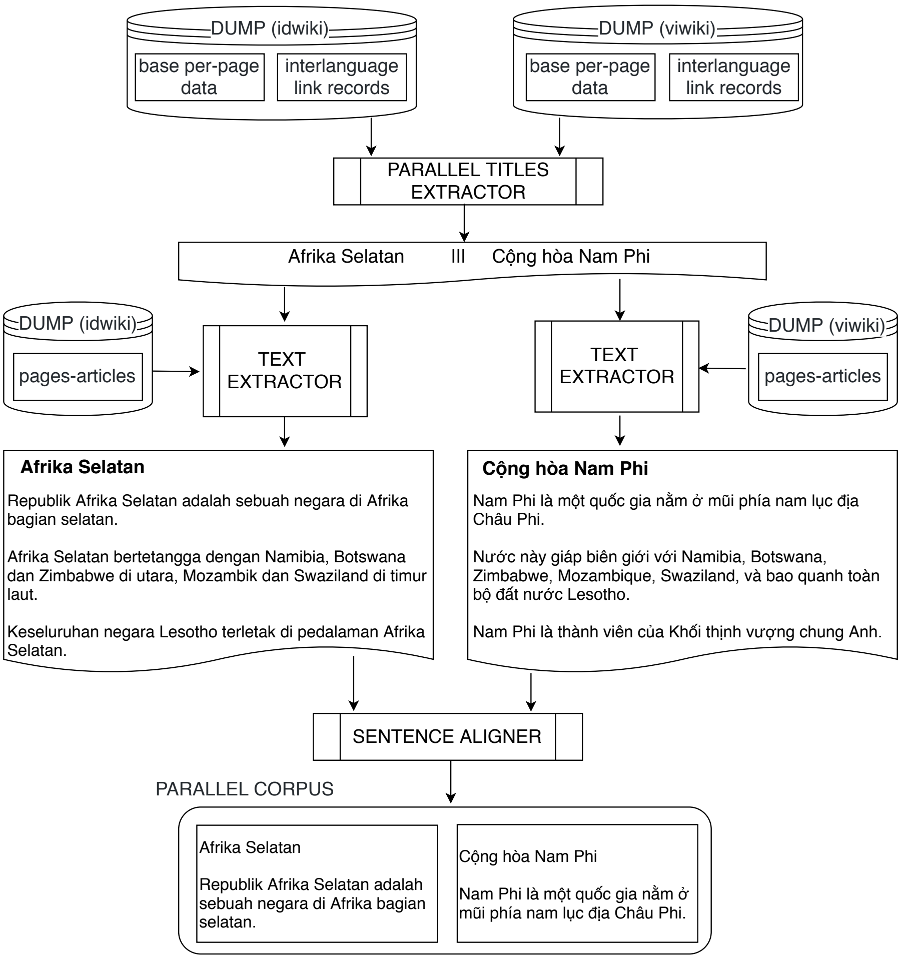
3.1 Parallel Title Extraction

In order to extract parallel titles of Wikipedia articles, we employed the resources from the Wikipedia dump.[10] Specifically, we employed the two following resources.

- **LANGIDwiki-TIME-page.sql.gz**: this is the base per-page data, which contains the page information such as id, title, old restrictions, etc.
- **LANGIDwiki-TIME-langlinks.sql.gz**: this resource contains the information of Wiki interlanguage link records.

[9] https://sites.google.com/a/uit.edu.vn/hungnq/evbcorpus

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



Afrika Selatan

Republik Afrika Selatan adalah sebuah negara di Afrika bagian selatan.

Afrika Selatan bertetangga dengan Namibia, Botswana dan Zimbabwe di utara, Mozambik dan Swaziland di timur laut.

Keseluruhan negara Lesotho terletak di pedalaman Afrika Selatan.

Cộng hòa Nam Phi

Nam Phi là một quốc gia nằm ở mũi phía nam lục địa Châu Phi.

Nước này giáp biên giới với Namibia, Botswana, Zimbabwe, Mozambique, Swaziland, và bao quanh toàn bộ đất nước Lesotho.

Nam Phi là thành viên của Khối thịnh vượng chung Anh.

Afrika Selatan

Republik Afrika Selatan adalah sebuah negara di Afrika bagian selatan.

Cộng hòa Nam Phi

Nam Phi là một quốc gia nằm ở mũi phía nam lục địa Châu Phi.

Fig. 1 Our model framework. An illustration on Indonesian-Vietnamese. *Afrika Selatan* (Indonesian), *Cộng hòa Nam Phi* (Vietnamese), English meaning: *South Africa*

where **LANGID** is the corresponding language codes: *en* (English), *vi* (Vietnamese), *id* (Indonesian), *ms* (Malay), and *tl* (Filipino).[11] **TIME** indicates the released time version. In this work, we used the database version *20200101*, which indicates the release time in 2020-01-01.

3.2 Article Extraction

Given a list of parallel titles for each language pair, we extract the corresponding articles' texts. The Wikipedia dump database provides us the available resource to extract articles' text.

[11] Tagalog is a language used in Philippines. Since there is no database available with the code *fil* (Filipino), we used the database of the code *tl* (Tagalog) for the Filipino in this work.

- **LANGIDwiki-TIME-pages-articles.xml.bz2**: the Wikipedia dump file containing article texts.

where **LANGID** is the language codes; **TIME** is the released time version.

The texts are then pre-processed including sentence split and word tokenization. We utilized the commonly used Moses scripts for both sentence split[12] and word tokenization.[13] Some languages may require language-specific word tokenizers such as Japanese, Thai, or Vietnamese, etc to obtain a better word tokenization quality. The Moses scripts may be a general and acceptable for some languages to some extent. We leave the detail language-specific processing for future work. We conduct a post-processing step after obtaining the final parallel corpora by detokenization[14] to return the original form of the extracted sentences.

Unbalance. In our observation, for some pair of articles, one article may contain hundreds of sentences but the other article only contains several sentences. It causes the sentence alignment task becomes more challenge and also time consuming. Therefore, we conduct a hard processing by keeping only the first l_s and l_t sentences for an article pair (S and T) so that the length (number of sentences) of the larger article (l_s) should be at most double the length of the other article (l_t), or if $l_s >> 2 * l_t$ then keep the first $l_s = 2 * l_t$ sentences.

3.3 Sentence Alignment

The last step is to extract parallel sentence pairs from each of the extracted parallel article pairs, or sentence alignment.

For the sentence alignment task, some powerful methods have been proposed such as the hunalign (Varga et al, 2005),[15] JMaxAlign (Kaufmann, 2012),[16] Bleualign (Sennrich and Volk, 2010),[17] and Microsoft Bilingual Sentence Aligner (Malign) (Moore, 2002).[18] In those systems, the Bleualign is a machine translation based algorithm, which requires an existing machine translation system and depends heavily on the translation provided (Sennrich and Volk, 2010). Meanwhile, the JMaxAlign is based on maximum entropy classifiers, which depend on a set of features such as length ratio, percentage of unaligned words, etc. The hunalign is one of the most well-known sentence alignment tools (Kaufmann, 2012). The two models hunalign and Malign are similar to some extent when they first extract parallel sentences based on sentence length ratio to build word alignment before extracting the final parallel sentences based on a combination of length-based and the word alignment. The hunalign is showed to be better in recall but worse in precision in comparison with the Malign (Varga et al, 2005). In this work, we choose Malign to extract parallel sentences from

[12] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl

[13] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[14] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/detokenizer.perl

[15] https://github.com/danielvarga/hunalign

[16] https://code.google.com/archive/p/jmaxalign/

[17] https://github.com/rsennrich/Bleualign

[18] https://download.microsoft.com/download/8/F/7/8F7F7CFF-46A2-40E0-80AC-2C3C3274C67B/bilingual-sentence-aligner.tar.gz

Table 1 Statistics of the Wikipedia database input

Data	Title		Text
	base per-page (MB)	interlanguage link (MB)	pages-articles (GB)
English	1800	373	16.8
Vietnamese	384	148	0.67
Indonesian	83	102	0.56
Malaysian	26	74	0.22
Filipino	7	31	0.05

Wikipedia articles; however, building the corpus based on the hunalign is also an interesting direction in future work.

For Malign (Moore, 2002), the model includes three phases. Firstly, parallel sentence pairs are extracted using a length-based method, in which the author assumed that the source and target sentence length are varied according to a Poisson distribution. Secondly, given the parallel sentence pairs extracted from the length-based phase, a word translation model is built using the well-known IBM Translation Model 1 (Brown et al, 1993). Finally, the parallel corpus is build based on a combination of the length-based and the word translation model. Since the model is a combination of both length-based and word translation, it is more accurate than the length-based model only. In addition, since the word translation model is built based on the sentence pairs extracted from the first phase, it does not depend on existing parallel resource to build the word model. Therefore, Malign can be suitable for our work in building parallel corpora from bilingual Wikipedia articles, and aiming at low-resource language pairs with unavailability of existing parallel corpora.

4 MT-WIKI Corpus

We present applying the method to extract parallel corpora for several low-resource South East Asian languages and between these languages paired with English namedly the *MT-WIKI Corpus*.

4.1 Wikidumps data

The input data for extracting the corpus is the Wikipedia database dumps. We used the released version **2020-01-01**. Table 1 presents the detail data size of the input resources: the base per-page data and the interlanguage link records.

The English data is in the largest size in both of the base per-page data and the interlanguage link records, and shows the much higher size than the other languages' data. Meanwhile, the Filipino data contains a limited amount. The data sizes imply that the parallel corpora of the languages paired with English are likely to be larger. Also, it is difficult to obtain large corpora of with such small input data as Filipino.

4.2 Extracted Parallel Titles and Articles

Table 2 presents the extracted parallel titles and articles from the Wikipedia dumps. It is noted that given a list of parallel titles, not all of the corresponding articles can be extracted.

Table 2 Statistics of extracted parallel titles and articles.

No.	Corpus	# Title	# Article
1	English-Vietnamese	334k	286k
2	English-Indonesian	289k	251k
3	English-Malay	240k	224k
4	Indonesian-Vietnamese	179k	158k
5	Malay-Vietnamese	137k	130k
6	Indonesian-Malay	129k	92k
7	English-Filipino	66k	58k
8	Filipino-Indonesian	44k	31k
9	Filipino-Vietnamese	39k	35k
10	Filipino-Malay	36k	27k

Table 3 Statistics of sentences in collected articles

No.	Corpus	# Source Sentences	# Target Sentences
1	English-Indonesian	16.5m	4.1m
2	English-Vietnamese	9.2m	3.4m
3	English-Malay	9.3m	2.0m
4	Indonesian-Vietnamese	2.2m	2.8m
5	Indonesia-Malay	1.6m	1.2m
6	Malay-Vietnamese	1.1m	1.9m
7	English-Filipino	4.4m	465k
8	Filipino-Indonesian	720k	293k
9	Filipino-Malay	396k	226k
10	Filipino-Vietnamese	292k	984k

The reason is that some articles are not exist or cannot be extracted given the titles because of some reasons such as encoding errors, the titles are for categories not for articles, etc. As a result, the number of extracted parallel articles is lower than the number of parallel titles.

In the comparison among the language pairs, the corpora of the languages paired with English are in the top (higher than 200k) (except for English-Filipino) because of the database dumps input size as we discussed. At the middle is the corpora between Indonesian, Malay, and Vietnamese (100-200k). The extracted parallel titles and articles for Filipino are in the lowest sizes with less than 100k titles and articles.

After obtaining parallel articles, we conduct several text processing including sentence split and word tokenization. Table 3 presents the statistics of sentences for each language pair as the input for extracting parallel sentence pairs. The statistics show that there is the unbalance between the number of sentences in the source and target sides some language pairs such as English-Indonesian (16.5 vs. 4.1 million of sentences), English-Malay (9.3 vs. 2.0 million of sentences). The corpora between Indonesian, Malay, and Vietnamese are somehow more balanced.

4.3 MT-WIKI Corpus

After conducting sentence alignment using the Malign (Moore, 2002), we obtain the final parallel corpora as presented in Table 4. In total, more than 2.6 million parallel sentences of ten language pairs between the low-resource South East Asian languages and the languages paired with English have been extracted.

Table 4 Statistics of the extracted parallel corpus (MT-WIKI)

#	Corpus	# Parallel Sentences	Average Source	Average Target
1	English-Vietnamese	865,100	14	17
2	English-Malay	534,317	13	11
3	English-Indonesian	384,135	16	14
4	Indonesian-Vietnamese	278,767	8	10
5	Malay-Vietnamese	225,162	8	10
6	Indonesian-Malay	102,964	12	12
7	English-Filipino	101,355	11	11
8	Filipino-Indonesian	46,021	6	8
9	Filipino-Vietnamese	44,233	6	6
10	Filipino-Malay	37,961	5	6
TOTAL		2,620,015		

On the top, the English-Vietnamese corpus is the largest data with more than 800k parallel sentences with the average of sentence length is 14 and 17 words per sentence, which is reasonably good for training translation models. The corpora of English-Malay, English-Indonesian, Indonesian-Vietnamese, and Malay-Vietnamese are also promising to train translation models when the data sizes are from 200k to 500k parallel sentences.

Meanwhile, as we discussed, with the limited database dumps input data, the corpora of the languages paired with Filipino are still quite small, and also contains short sentences with the average sentence length is less than 10 words per sentence. It may be difficult to obtain good translation performance on such limited corpora.

5 Evaluation on Machine Translation

We conduct evaluation by using the parallel corpora for training machine translation models since one of our main goals is to build parallel corpora for improving machine translation on low-resource languages. In this section, we present the experiments and results to evaluate the extracted corpus for machine translation task.

5.1 Experimental Settings

5.1.1 Dataset

We conducted experiments on two tasks: the Asian Language Treebank (Riza et al, 2016) (ALT) and the IWSLT 2015 (Cettolo et al, 2015) data sets, which are the existing corpora for the low-resource Southeast Asian languages. Both data sets are manually translated, which are reliable for the evaluation. Additionally, both data sets are publicly available and widely used, so we believe that conducting evaluation on these tasks are easier for comparison or reproduction in future work.

- **Asian Language Treebank** (ALT) (Riza et al, 2016)[19] – This is a parallel treebank for ten languages including English, Filipino, Indonesian, Malay, Vietnamese, Japanese, and some other Asian languages. The corpus is built from 20,000 English sentences and

[19] `http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/`

manually translated into the other languages by native speakers. This corpus aims to advance the state-of-the-art Asian natural language processing techniques. To our best knowledge, our experiments is the first evaluation of this corpus on machine translation task. Fortunately, the authors divided the data into training, development, and test sets of 18k, 1k, and 1018 parallel sentences, respectively. Therefore, we directly used the official data split for our experiments.

- **IWSLT** (Cettolo et al, 2012) – This is a spoken language data set which contains subtitles in TED Talks, which are used for the well-known IWSLT evaluation campaigns (Cettolo et al, 2015). We conducted experiments on the IWSLT 2015 evaluation campaign[20] containing the English-Vietnamese translation. The task provided 129k parallel sentences for training data. Following some previous work (Luong and Manning, 2015; Clark et al, 2018), we used the same data, which is the officially provided *tst2012* data set for the development set (1,553 sentences), and *tst2013* for test set (1,268 sentences).
- **MT-WIKI**: our multilingual parallel corpus on the ten language pairs of English, Filipino, Indonesian, Malay, and Vietnamese extracted from Wikipedia texts.

For the ALT task, we conducted the evaluation on ten language pairs from English, Filipino, Indonesian, Malay, and Vietnamese on two translation directions. In total, we created twenty translation models. It is noted that the IWSLT task is only available for English-Vietnamese.

We conducted preprocessing the data sets on word tokenization, cleaning (remove sentences containing more than 175 words), and lowercase using the commonly used mosesdecoder scripts (Koehn et al, 2007).[21] For processing to train neural-based models, we applied subword segmentation using byte pair encoding (Sennrich et al, 2016).

5.1.2 Training Settings

Transformer. In order to train translation models, we choose one of the best current methods, the Transformer (Vaswani et al, 2017; Edunov et al, 2018). Specifically, we employed the recent *fairseq* model developed by Facebook AI (Ott et al, 2019) implemented on Pytorch.[22]

For the training parameters, we used the basic provided $transformer$ architecture ("$-arch$" option). This architecture is the base model with the number of encoder layers $N = 6$, encoder embedding dimension $d_{model} = 512$, number of attention heads $h = 8$. The model is trained with the Adam optimizer Kingma and Ba (2015), in which $adam - betas = (0.9, 0.98)$; $dropout = 0.1$; weight decay as 0.0001; and max tokens 4096.

For evaluation, results are generated use beam search with a beam width of 5 and the batch size 128. The scores are reported based on the $BLEU4$ from the multi-bleu available in the *fairseq*.

5.1.3 Compared Models

For each of the two tasks ALT and IWSLT, we built baseline models using the ALT and IWSLT training data sets. In order to evaluate the contribution of the *MT-WIKI* corpus, we

[20] `https://sites.google.com/site/iwsltevaluation2015/mt-track`

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

[22] `https://github.com/pytorch/fairseq`

created two different settings: using the *MT-WIKI* corpus only and merge with the baseline training data. All of the models are trained using the same training setting as we described in Section 5.1.2. In particular, we describe the models as follows.

- **Base_ALT**: The models are trained on the baseline ALT training sets.
- **WikiSingle_ALT**: The models are trained on the *MT-WIKI* corpus.
- **WikiEnhanced_ALT**: the models are trained on the *MT-WIKI* corpus merged with the baseline ALT training sets.

Similarly, we created three settings for the IWSLT task.

- **Base_IWSLT**: The models are trained on the baseline IWSLT training set.
- **WikiSingle_IWSLT**: The models are trained on the *MT-WIKI* English-Vietnamese corpus.
- **WikiEnhanced_IWSLT**: The models are trained on the *MT-WIKI* English-Vietnamese corpus merged with the baseline IWSLT training set.

In addition, we compared with several previous work evaluated on the IWSLT task.

- **TensorNMT**[23]: A tutorial of neural machine translation released by the well-known Tensorflow, in which they reported the results on the IWSLT 2015 task.
- **Stanford** (Luong and Manning, 2015): This is a neural-based system participated in the IWSLT 2015 task, which is LSTM networks (Hochreiter and Schmidhuber, 1997) of 4 layers and attention mechanism (Luong et al, 2015).
- **CVT**(Clark et al, 2018): A semi-supervised learning algorithm improved from bidirectional LSTM.

5.2 Results

5.2.1 Results on ALT Task

Table 5 presents the performance on the ALT test set. In overall, although using only the *MT-WIKI* corpus, there are 9/20 models obtaining the better performance than the baseline models. When we combine the baseline and the *MT-WIKI* corpus for training, we achieve the significant improvement on most language pairs (except for the corpora of Filipino).

With the small corpus size on the Filipino corpora (less than 50k parallel sentences), it is difficult to train translation models, which leads to the low performance. However, for English-Filipino with a bit higher size (101k parallel sentences), we can obtain a better performance on English-Filipino when training the *MT-WIKI* corpus only, and obtain the improvement on both directions when we combine the *MT-WIKI* corpus with the baseline data.

5.2.2 Results on IWSLT Task

Results on IWSLT task are presented in Table 6. The models trained on only the *MT-WIKI* corpus obtained reasonably high results and even better than the baseline on the English-Vietnamese *tst2013* test set. When combining the baseline data with the *MT-WIKI* corpus, we achieved the best performance on all cases and significantly improved the baseline from 3 to 4 BLEU points.

[23] `https://github.com/tensorflow/nmt`

Table 5 Results on the ALT test set.

Pair	**Model**		
	Base_ALT	**WikiSingle_ALT**	**WikiEnhanced_ALT** (BLEU)
English-Filippino	17.77	**19.16**	**28.74**
English-Indonesian	25.57	**37.94**	**40.18**
English-Malay	30.57	**42.42**	**45.61**
English-Vietnamese	20.37	**39.10**	**39.82**
Filippino-English	22.10	15.02	**28.58**
Filippino-Indonesian	8.96	0.22	8.93
Filippino-Malay	10.13	0.20	7.41
Filippino-Vietnamese	7.07	0.37	4.28
Indonesian-English	24.74	**35.02**	**37.87**
Indonesian-Filippino	10.57	0.86	**10.66**
Indonesian-Malay	26.13	25.34	**31.93**
Indonesian-Vietnamese	8.52	7.71	**17.55**
Malay-English	27.02	**37.71**	**41.11**
Malay-Fillippino	11.14	0.76	9.45
Malay-Indonesian	27.57	**27.57**	**32.71**
Malay-Vietnamese	10.36	6.25	**15.92**
Vietnamese-English	11.47	**30.81**	**33.29**
Vietnamese-Fillipino	8.91	1.09	4.25
Vietnamese-Indonesian	7.84	**8.79**	**17.27**
Vietnamese-Malay	9.42	7.42	**18.12**

Table 6 Results on the IWSLT test set.

Model	**English-Vietnamese**		**Vietnamese-English**	
	tst2012	tst2013	tst2012	tst2013 (BLEU)
TensorNMT	23.8	26.1	–	–
Stanford (Luong and Manning, 2015)	–	26.9	–	–
CVT (Clark et al, 2018)	–	29.6	–	–
Base_IWSLT	26.26	29.58	24.44	27.36
WikiSingle_IWSLT	25.14	**30.36**	19.24	23.59
WikiEnhanced_IWSLT	**29.60**	**33.58**	**27.95**	**31.79**

In comparison with the previous work, the baseline performance is comparable with the CVT (Clark et al, 2018) (29.6 BLEU) on the *tst2013*. When using the *MT-WIKI* corpus, we obtained the improvement of +0.8 and 4 BLEU points on the settings of using the *MT-WIKI* corpus only and the combined data, respectively.

On the English-Vietnamese *tst2012* data set, the baseline is 2.4 BLEU points higher than the TensorNMT. Meanwhile, using the *MT-WIKI* corpus combined with the baseline data helps to improve 3.3 BLEU points.

5.3 Discussion

5.3.1 Wikidump Data.

The size of extracted corpora depends on the size of available Wikidump data. For several language pairs such as English-Vietnamese, English-Malay, etc, since the Wikidump data is available with a larger size, we can build larger parallel corpora and obtain better translation performance, which is competitive or even better than the existing manually created corpus. However, when the Wikidump data is quite small as the case of the Filipino corpora, it is difficult to obtain such improvement. The issue can be partly solved in the future when Wikipedia articles are increasingly added.

5.3.2 Low-Resource Languages.

In this work, we evaluated the translation performance on some yet investigated low-resource language pairs such as Indonesian-Vietnamese, Malay-Filipino, Vietnamese-Malay etc, and obtained some first results on these language pairs. From the observation in Table 5, the results on these language pairs are still limitation, for instance Indonesian-Vietnamese (8.52 BLEU), Malay-Filipino (11.14 BLEU), or Vietnamese-Malay (9.42 BLEU). Further investigation on these specific low-resource languages may need to be conducted in future research.

5.3.3 Translation Qualitative.

We present some instances extracted from our models' translation output.

6 Conclusion

In this work, we introduce a multilingual parallel corpus of 2.5 million parallel sentences on ten language pairs of Southeast Asian languages including Filipino, Indonesian, Malay, Vietnamese, and these languages paired with English. The motivation comes from the fact that there are few or even no existing parallel corpora on such languages, which are an essential resource in building machine translation. The corpus is built from the abundantly available Wikipedia resource on two levels. On the document level, parallel articles are extracted by utilizing the available inter-language link information on the Wikipedia dumps. On the sentence level, parallel sentences are aligned based on an existing powerful and language-independent sentence aligner. We evaluated the corpus on machine translation on two tasks: the Asian Language Treebank corpus and the IWSLT 2015 shared task. Translation systems are built using the state-of-the-art Transformer model. Experimental results on various settings showed that our corpus significantly improve the translation performance on these low-resource languages. The corpus and code are released for research community, which can help to enhance the research on machine translation on such low-resource languages. The framework can be applied to build parallel corpora for other languages. In future work, we plan to improve the sentence level alignment such as trying different aligners or extracting shorter parallel units instead of sentences such as parallel fragments or clauses, which may be more suitable on such noisy and comparable Wikipedia articles.

References

Adafre SF, De Rijke M (2006) Finding similar sentences across multiple languages in wikipedia. In: Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources

Barrón-Cedeño A, España-Bonet C, Boldoba J, Màrquez L (2015) A factory of comparable corpora from wikipedia. In: Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pp 3–13

Table 7 Translation samples.

Language	Model	Sample	Meaning (English)
id-vi	Input(Indonesian)	dia mengatakan bahwa dia tidak yakin berapa banyak penerbangan yang terpengaruh karena letusan tersebut.	he said that he was uncertain of how many flights were affected due to the eruption.
	Reference(Vietnamese)	ông nói mình không chắc có bao nhiêu chuyến bay bị ảnh hưởng vì vụ phun trào.	
	Base_ALT	ông nói rằng ông không có tin rằng không có ảnh hưởng đến các chuyến bay bị ảnh hưởng bởi sự kiện này.	**he said** he does not believe that no **influence** on **flights was affected by** the event.
	WikiEnhanced_ALT	ông nói rằng ông đã không tin có bao gồm nhiều chuyến bay bị ảnh hưởng bởi vụ phun trào.	**he said** he did not believe that many **flights were affected by the eruption.**
en-ms	Input(English)	four of the children were ejected from the truck and died at the scene.	four of the children were ejected from the truck and died at the scene.
	Reference(Malay)	empat kanak-kanak telah dikeluarkan dari trak itu dan meninggal dunia di tempat kejadian.	
	Base_ALT	empat dari trak tersebut disuntik dari trak tersebut dan meninggal dunia di tempat kejadian.	**four of the** truck were injected **from the truck and died at the scene.**
	WikiSingle_ALT	empat daripada anak-anak telah dikeluarkan dari lori dan meninggal dunia di tempat kejadian.	**four of the children were** released **from the truck and died at the scene.**
vi-en	Input(Vietnamese)	bài thuyết trình nào bạn vỗ tay nhiều nhất trong sáng nay?	which presentation have you applauded the most this morning?
	Reference(English)	which presentation have you applauded the most this morning?	
	Base_IWSLT	what are the most powerful hand you can clapping in this morning?	
	WikiEnhanced_IWSLT	what is the most clapping presentation in the morning?	

Bojar O, Dušek O, Kocmi T, Libovickỳ J, Novák M, Popel M, Sudarikov R, Variš D (2016) Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered. In: International Conference on Text, Speech, and Dialogue, Springer, pp 231–238

Brown PF, Pietra VJD, Pietra SAD, Mercer RL (1993) The mathematics of statistical machine translation: Parameter estimation. Computational linguistics 19(2):263–311

Cettolo M, Girardi C, Federico M (2012) Wit^3: Web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT), Trento, Italy, pp 261–268

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

Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp 1724–1734, DOI 10.3115/v1/D14-1179, URL `https://www.aclweb.org/anthology/D14-1179`

Chu C, Nakazawa T, 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, DOI 10.1145/2833089, URL `http://doi.acm.org/10.1145/2833089`

Clark K, Luong MT, Manning CD, Le Q (2018) Semi-supervised sequence modeling with cross-view training. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 1914–1925

Edunov S, Ott M, Auli M, Grangier D (2018) Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 489–500

Gupta R, Pal S, Bandyopadhyay S (2013) Improving mt system using extracted parallel fragments of text from comparable corpora. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, pp 69–76

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780

Kaufmann M (2012) Jmaxalign: A maximum entropy parallel sentence alignment tool. In: Proceedings of COLING 2012: Demonstration Papers, pp 277–288

Kingma DP, Ba JL (2015) Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR2015)

Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the Tenth Machine Translation Summit (MT Summit X), Phuket, Thailand, URL `http://mt-archive.info/MTS-2005-Koehn.pdf`

Koehn P, Och FJ, 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, Association for Computational Linguistics, pp 48–54

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, Association for Computational Linguistics, pp 177–180

Luong MT, Manning CD (2015) Stanford neural machine translation systems for spoken language domains. In: Proceedings of the International Workshop on Spoken Language Translation, pp 76–79

Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp 1412–1421

Mohammadi M, GhasemAghaee N (2010) Building bilingual parallel corpora based on wikipedia. In: 2010 Second International Conference on Computer Engineering and Applications, IEEE, vol 2, pp 264–268

Moore RC (2002) Fast and accurate sentence alignment of bilingual corpora. In: Conference of the Association for Machine Translation in the Americas, Springer, pp 135–144

Ng N, Yee K, Baevski A, Ott M, Auli M, Edunov S (2019) Facebook fair's wmt19 news translation task submission. In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pp 314–319

Ngo QH, Winiwarter W, 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), pp 1–9

Otero PG, Lopez IG (2010) Wikipedia as multilingual source of comparable corpora. In: Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC, pp 21–25

Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, Grangier D, Auli M (2019) fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp 48–53

Plamadă M, Volk M (2013) Mining for domain-specific parallel text from Wikipedia. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Association for Computational Linguistics, Sofia, Bulgaria, pp 112–120, URL `https://www.aclweb.org/anthology/W13-2514`

Resnik P (1999) Mining the web for bilingual text. In: Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL), URL `http://acl.ldc.upenn.edu/P/P99/P99-1068.pdf`

Riza H, Purwoadi M, Uliniansyah T, Ti AA, Aljunied SM, Mai LC, Thang VT, Thai NP, Chea V, Sam S, et al (2016) Introduction of the asian language treebank. In: 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, pp 1–6

Saad M, Langlois D, Smaïli K (2013) Extracting comparable articles from wikipedia and measuring their comparabilities. Procedia-Social and Behavioral Sciences 95:40–47

Sennrich R, Volk M (2010) Mt-based sentence alignment for ocr-generated parallel texts. In: Proceedings of AMTA 2010

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 1715–1725

Singh S, Kunchukuttan A, Bhattacharyya P (2016) Iit bombay's english-indonesian submission at wat: Integrating neural language models with smt. In: Proceedings of the 3rd Workshop on Asian Translation (WAT2016), pp 68–74

Smith JR, Quirk C, Toutanova K (2010) Extracting parallel sentences from comparable corpora using document level alignment. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp 403–411

Ştefănescu D, 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), pp 24–30

Steinberger R, Pouliquen B, Widiger A, Ignat C, Erjavec T, Tufiş D, Varga D (2006) The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)

Tan L (2016) Faster and lighter phrase-based machine translation baseline. In: Proceedings of the 3rd Workshop on Asian Translation (WAT2016), The COLING 2016 Organizing Committee, Osaka, Japan, pp 184–193

Tan L, Bond F (2011) Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). In: Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, pp 362–371

Tian L, Wong DF, Chao LS, Quaresma P, Oliveira F, Yi L (2014) Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In: LREC, pp 1837–1842

Tiedemann J (2012) Parallel data, tools and interfaces in opus. In: Lrec, vol 2012, pp 2214–2218

Utiyama M, Isahara H (2007) A japanese-english patent parallel corpus. Proceedings of MT summit XI pp 475–482

Varga D, Halácsy P, Kornai A, Nagy V, Németh L, Trón V (2005) Parallel corpora for medium density languages. In: Proceedings of the RANLP 2005, pp 590–596

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008

Wang P, Nakov P, Ng HT (2016) Source language adaptation approaches for resource-poor machine translation. Computational Linguistics 42(2):277–306

Wang Y, Xia Y, He T, Tian F, Qin T, Zhai CX, Liu TY (2019) Multi-agent dual learning. In: 7th International Conference on Learning Representations, ICLR 2019

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

Ziemski M, Junczys-Dowmunt M, Pouliquen B (2016) The united nations parallel corpus v1. 0. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp 3530–3534