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

HAI-LONG TRIEU, Vingroup Big Data Institute, Hanoi, Vietnam
ASHWIN ITTOO, University of Liège, Belgium

Parallel corpora play an essential role in machine translation (MT) task. High quality MT systems depend mainly on the availability of large scale parallel corpora. However, such large parallel corpora are only available for several languages such as English, German, French, Chinese, Japanese, or Arabic, etc. Most other languages including Southeast Asian area are low-resource, in which such large corpora are unavailable. This issue limits the development of MT task in these languages. In this work , in order to make a step to fill this gap, we introduce a multilingual parallel corpus for Southeast Asian languages containing more than 2.6 million parallel sentences ranging in ten language pairs of the Southeast Asian (SeA) languages including Indonesian, Malay, Filipino, Vietnamese and these languages paired with English. The corpus is automatically extracted by utilizing the available abundant Wikipedia resources. In order to build the corpus, our methods consists of three main steps. First, parallel titles of Wikipedia articles are extracted from the available Wikipedia interlanguagelink data. Then, parallel articles' texts are collected. Finally, parallel sentences are aligned to construct the corpus. In order to evaluate the contribution of our corpus in improving MT task, we conducted experiments on the two benchmark datasets, i.e., the Asian Language Treebank and the IWSLT shared task. The results confirm the contribution of our corpus when MT systems trained on our corpus achieve promising results and improvement in both scenarios, i.e., using our corpus only or combining with existing data. This corpus can advance the development of MT task in the SeA languages. The code and data are publicly available at https://github.com/longth-vbd/sea-wiki-parl.

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

Additional Key Words and Phrases: parallel corpus, multilingual corpus, machine translation, Southeast Asian languages

ACM Reference Format:
Hai-Long Trieu and Ashwin Ittoo. 2022. A Multilingual Parallel Corpus for Improving Machine Translation on Southeast Asian Languages. 1, 1 (May 2022), 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large parallel corpora of translated sentence pairs are a critical resource at the core of machine translation (MT) systems [7, 17, 43]. These corpora exist predominantly for European language pairs, such as English-German, English-French and Czech-English [3, 16, 36]. With regards to Asian languages, research has been primarily targeted at Chinese (English-Chinese) [39] and Japanese (English-Japanese) [41]. We refer to these language pairs as "resource-rich". The availability of such large, parallel corpora have enabled significant development and state-of-the-art performance in MT in these resource-rich language pairs [11, 24, 46].

Authors' addresses: Hai-Long Trieu, Vingroup Big Data Institute, Hanoi, Hanoi, Vietnam, v.longth12@vinbigdata.com; Ashwin Ittoo, University of Liège, Belgium, Ashwin.Ittoo@uliege.be.

Manuscript submitted to ACM

However, MT research on several other languages, and in particular, a subset of Southeast Asian (SeA) languages, such as Filipino, Indonesian, Malay, Vietnamese, has been largely overlooked. We refer to them as "low-resource" (or "resource-poor") languages. Research in these languages has not experienced the same success as that of their resource-rich counterparts, and is still lagging behind. As can be expected, 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. Existing parallel corpora are scare and of so small size that they do not constitute a reliable and statistically-significant source of training data [30, 38]. Subsequently, the performance of MT systems trained on such data for SeA language translation fares poorly [33, 37, 45].

In terms of languages, it is noted that SeA languages are prominently used by large populations [47]. For instance, Malay and Vietnamese are used by more than 200 millions people, and in the top around 14th-16th languages used in the world. With the rapid development of this SeA area recently [44], there is a high demand for the exchange among these countries and languages, and MT should play a key role. In order to enable further developments in MT research across SeA low-resource languages, it is therefore urgent to alleviate the issue of insufficient training data, which is fundamental to improve MT performance.

In this article, we make a first, significant step towards in this direction. Specifically, as our main contribution, we develop and present a large multilingual parallel corpus, with more than 2.6 million sentence pairs, encompassing four SeA languages, viz. Indonesian, Malay, Filipino and Vietnamese, all paired with English. Our methodology for constructing the parallel corpus is fully automated, thereby overcoming the traditional challenges of manual corpus creation, which is often tedious and time-consuming. Our methodology relies on Wikipedia dumps, a publicly available repository containing a large number of article texts written in various languages including the SeA languages. 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.

To evaluate the quality of our automatically created parallel corpus, we compared the performance of a Transfomer-based MT system [27] trained on it against the performance when the MT system is trained on two manually constructed corpora, viz. the Asian Language Treebank (ALT) corpus [30] and the IWSLT 2015 shared tasks [6]. Our experiments results revealed that the performance of the MT system, when trained solely on the manually created corpora, improved significantly when our automatically created corpus was incorporated during training. In addition, the performance achieved when training on our automatically created corpus was comparable to that achieved when training on the manually created ones, which as noted above, could be reasonably expected to be of higher quality. These promising results suggest that our proposed parallel corpus is of comparable quality to manually created ones. This despite it being created automatically, devoid of (or with very minimal) human intervention. These results also highlight the reliability of our proposed (automated) methodology for constructing parallel corpus. As already noted, the methodology can be straightforwardly extended to other language pairs. It alleviates the challenges involved in manual corpus creation, and can help the development of MT research across other language pairs which have been overlooked todate.

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 potentially improve other multilingual NLP tasks on these languages.
(2) We conducted intensive experiments and analyses to evaluate our corpus for MT on the SeA languages. Several results that we evaluated on the standard corpora can be served as a benchmark for these low-resource yet

investigated languages in MT tasks. Additionally, MT systems on low-resource languages, which lack available parallel data, can apply our strategy of utilizing the abundant Wikipedia resources to improve the performance.

(3) In addition, our methodology is simple, but yet very promising 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, especially other low-resource languages, given such Wikipedia dumps resources, which are abundant and available for most languages. We release the corpus and code, which is publicly available at https://github.com/longth-vbd/sea-wiki-parl.

2 RELATED WORK

Large parallel corpora play an essential role in cross-lingual natural language processing tasks including machine translation. Several multilingual corpora have been collected and built including the *Europarl* [16],[1] *JRC-Acquis* [36],[2] *UN Parallel Corpus* [48],[3] *WIT3* corpus [5],[4] and *OPUS* [40].[5] 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. For Asian languages, there are a few large bilingual corpora such as the *UM-Corpus* English-Chinese [39][6] collected from bilingual websites, and the *NTCIR PatentMT* corpus[7] of Japanese-English and Chinese-English collected from patent description. These corpora are large and reliable with high quality. However, they are limited in only several rich resource languages and depend on the availability of the resources.

An alternative strategy is to automatically build and augment parallel corpora from online resources [29], such as Wikipedia [1, 12, 34], a large publicly available articles on many languages. Many methods have been proposed to extract parallel texts from Wikpedia such as based on sentence similarity [1], linked-based method [22], binary and cosine similarity [31], cross-lingual information retrieval [35], document level alignment with maximum entropy classifier [34], clause level alignment [28], or extracting parallel fragments [9, 12]. However, most methods are applied on European languages such as English, Dutch, Spanish, Portuguese, German, Polish, Bulgarian [1, 2, 26, 35]. For Asian languages, several corpora are built based on Wikipedia such as Persian-English [26], English-Bengali [12], or Chinese-Japanese [9]. There is no prior work to build parallel corpus from Wikipedia for Southeast Asian languages, to our best knowledge.

For building parallel corpora on Southeast Asian languages, there are several efforts [25, 30, 38]. For *English-Vietnamese*, the *EVBCorpus* [25] contains 800K parallel sentences and its updated version[8] 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 [38] containing 15K sentences in six languages English, Chinese, Japanese, Korean, Indonesian, and Vietnamese and the *Asian Language Treebank (ALT)* corpus [30] 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

[1] https://www.statmt.org/europarl/
[2] https://ec.europa.eu/jrc/en/language-technologies
[3] https://conferences.unite.un.org/uncorpus/
[4] https://wit3.fbk.eu
[5] http://opus.nlpl.eu
[6] http://nlp2ct.cis.umac.mo/um-corpus/
[7] http://ntcir.nii.ac.jp/PatentMT/
[8] https://sites.google.com/a/uit.edu.vn/hungnq/evbcorpus

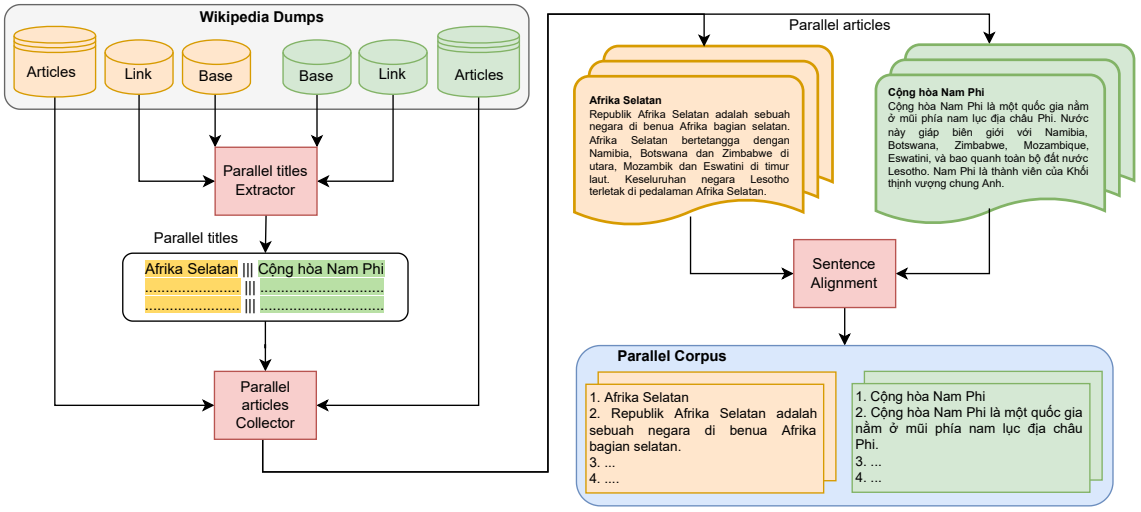

Fig. 1. Overview of our method to build the parallel corpus from Wikipedia. We illustrate an example of extracting parallel sentences for Indonesian-Vietnamese from the article pair collected from the parallel titles: *Afrika Selatan* (in Indonesian) paired with Cộng hòa Nam Phi (in Vietnamese) (meaning: *South Africa*). The Wikpedia Dumps resources are used as the input (*Base*: the base per-page data, *Link*: the interlanguage link data, and *Articles*: the pages articles data)

translating English texts into the other languages. Since most of these works are based on available parallel resources such as bilingual webs or documents or manually translated texts, the corpus sizes are limited. Lacking large parallel corpora leaves a gap for machine translation for the languages in Southeast Asian area.

3 METHOD

We present the method we used to build a parallel corpus from Wikipedia for Southeast Asian languages. The method includes three main steps: extracting parallel titles, collecting parallel articles, and aligning bilingual sentences. First, we extract parallel titles of Wikipedia articles available from Wikipedia dumps. Then, we collect corresponding parallel articles given the extracted parallel titles. Finally, sentences from each pair of parallel articles are aligned to extract the bilingual corpus. We illustrate the method overview in Figure 1.

3.1 Extracting Parallel Titles

Wikipedia resources. We extract parallel titles of Wikipedia articles by employing the resources from Wikipedia dump.[9] Specifically, we employed the two resources as the following patterns.

(1) **Base per-page data**: *{language_code}***wiki**-*{released_time}*-**page.sql.gz**: this resource contains the page information such as id, title, old restrictions, etc.

(2) **Interlanguage link data**: *{language_code}***wiki**-*{released_time}*-**langlinks.sql.gz**: this resource contains the information of Wiki interlanguage link records, which are links from a page in Wikipedia to an equivalent page in another language.

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

Table 1. Data sizes of the input Wikipedia resources (MB) to 1) extract parallel titles (Page data and Interlanguage data) and 2) collect articles' texts (Page articles data)

Language	Page data	Interlanguage data	Page articles data
English	1800	373	1,680
Vietnamese	384	148	670
Indonesian	83	102	560
Malaysian	26	74	220
Tagalog	7	31	50

Table 2. Extracting parallel titles (we present the languages in descending order of data sizes

	English	Vietnamese	Indonesian	Malay	Tagalog
English	–	334K	289K	240K	66K
Vietnamese	–	–	179K	137K	39K
Indonesian	–	–	–	129K	44K
Malay	–	–	–	–	36K

where *{language_code}* is the language code for a language such as *en* for English and *vi* for Vietnamese, etc. Meanwhile, *{released_time}* shows the timestamp that the resource is released.[10] For instance, in order to extract the bilingual corpus for English-Vietnamese, we need the following resources:

(1) **en**wiki-20200101-page.sql.gz
(2) **en**wiki-20200101-langlinks.sql.gz
(3) **vi**wiki-20200101-page.sql.gz
(4) **vi**wiki-20200101-langlinks.sql.gz

Selected languages. For the scope of this work, we selected several Southeast Asian languages in which the Wikipedia resources are available. In addition, we focus on the languages with Latin script. Non-Latin script languages such as Thai (Thai script), Lao (Lao script), or Burmese (for Myanmar, Burmese script) are left for future research. For the Philippines language, we found that Filipino Wikipedia resource unavailable, but Tagalog (language code: tl) resources are available, while Tagalog is one of the main languages used in Philippines. Alternatively, we thus used the Tagalog Wikipedia resources to extract the data for Philippines's language. As a result, we selected the following languages to build our corpus: Indonesian (language code: id), Malay (ms), Vietnamese (vi), Tagalog (tl), and English (en).

Extracted parallel titles. We downloaded the Wikipedia resources for the selected languages. Then, we extracted the parallel titles for each language pair.[11] We presented the input Wikipedia resources in Table 1, and extracted parallel titles in Table 2.

3.2 Collecting Parallel Articles

Page articles data. Given the extracted Wikipedia titles, we collect the corresponding texts for each article. We employed the Wikipedia resource containing articles' texts, which is in the following format.

- **Pages articles data**: *{language_code}-{released_time}*-**pages-articles.xml.bz2**

[10] In this work, we built our corpus based on the database version *20200101*, of which the resources are released in 2020-01-01.
[11] We used the script to extract parallel titles: https://github.com/clab/wikipedia-parallel-titles

Table 3. Collecting parallel articles.

	English	Vietnamese	Indonesian	Malay	Tagalog
English	–	286K	251K	224K	58K
Vietnamese	–	–	158K	130K	35K
Indonesian	–	–	–	92K	31K
Malay	–	–	–	–	27K

Table 4. The number of sentences in collected parallel articles

#	**Language pair**	**Source**	**Target**
1	English-Indonesian	16.5M	4.1M
2	English-Vietnamese	9.2M	3.4M
3	English-Malay	9.3M	2.0M
4	English-Tagalog	4.4M	465K
5	Indonesian-Vietnamese	2.2M	2.8M
6	Indonesia-Malay	1.6M	1.2M
7	Malay-Vietnamese	1.1M	1.9M
8	Tagalog-Indonesian	720K	293K
9	Tagalog-Malay	396K	226K
10	Tagalog-Vietnamese	292K	984K

where *{language_code}* is the language code of a language (en, vi, id, ms, tl), and *{released_time}* shows the released time of the resource.

Collected parallel articles. We present the results of collected parallel articles in Table 3.

3.3 Aligning Parallel Sentences

Given the extracted parallel articles, we align sentences in each article pair to extract parallel sentences. We first describe preprocessing steps on the collected articles. Then, we discuss sentence aligners and focus on the *Microsoft Bilingual Sentence Aligner* [23], which we used for aligning parallel sentences. Finally, we present the results of aligned parallel sentences.

Preprocessing. We conducted preprocessing steps including splitting sentences and word tokenization. For these tasks, we employed the commonly used Moses scripts for both sentence splitting[12] and word tokenization.[13] A language may have its specific preprocessing tools. However, in this work, we only apply the Moses tokenizers for all of the languages for the simplest setting and language independence. Applying specific preprocessing tools for each language may further improve the performance, and we leave this application for future work. We present the number of sentences in the collected articles after preprocessing in Table 4. This is the input for the next step, sentence alignment, presented below.

[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

Table 5. Our extracted multilingual parallel corpus. We show the number of parallel sentences in each language pair with corresponding average sentence length of the source and target languages.

#	Language pair	Parallel sentences	Source Length	Target Length
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-Tagalog	101,355	11	11
8	Tagalog-Indonesian	46,021	6	8
9	Tagalog-Vietnamese	44,233	6	6
10	Tagalog-Malay	37,961	5	6
	TOTAL	2,620,015		

Sentence aligners. Sentence alignment is an important task in building parallel corpora. Several methods for sentence alignment have been proposed in literature including *Hunalign* [42],[14] *JMaxAlign* [14],[15] *Bleualign* [32],[16] and *Microsoft Bilingual Sentence Aligner (Malign)* [23]. In these methods, *Bleualign* requires an existing machine translation system, which typically depends on the quality and availability of parallel corpora to train the translation model. In addition, its performance depends heavily on the translation quality provided [32]. *JMaxAlign* uses maximum entropy classifiers to classify parallel sentences. It depends on an external word aligner, Berkeley Aligner [18] to align words from parallel sentences and extract features for the classifier. *Hunalign* [42] and *Malign* [23] are the two well-known sentence aligners, which are language independent and do not depend on any external tools or existing parallel training data. Both methods build word alignment models based on sentence length-based in the first phase, and combining aligned words with sentence length in the second phase. According to the analyses of [42], *Hunalign* outperforms *Malign* in recall, but worse precision.

Microsoft Bilingual Sentence Aligner (Malign). We chose *Malign*, a well-known, language independent, and high precision sentence aligner to align parallel sentences for building our corpus from Wikipedia texts. *Malign* includes three main steps. First, a length-based method is used to extract the first-phase parallel sentence pairs following the Poisson distribution. Then, these parallel sentences are used to build a word alignment model based on the well-known IBM Translation Model 1 [4]. Finally, the length-based method is used combining with the word alignment model to extract parallel sentences.

Our extracted corpus. After aligning sentences using *Malign*, we achieved a multilingual parallel corpus between the Southeast Asian languages including Indonesian, Malay, Tagalog, and Vietnamese, and also these languages paired with English. Totally, we achieved more than 2.6 million parallel sentences of ten language pairs. The number of parallel sentences in each language pair is presented in Table 5. We named our corpus as *SeA-Wiki-Parl* which stands for **S**outh**e**ast **A**sian **Wiki**pedia-based **Par**al**l**el Corpus.

[14] https://github.com/danielvarga/hunalign
[15] https://code.google.com/archive/p/jmaxalign/
[16] https://github.com/rsennrich/Bleualign

Table 6. Datasets used in our experiments for MT task. We present the number of parallel sentences in the training, development, and test sets, and the languages in each corpus (*en*: English, *id*: Indonesian, *ms*: Malay, *tl*: Tagalog, *fil*: Filipino, and *vi*: Vietnamese)

Corpus	Train	Dev	Test	Languages
SeA-Wiki-Parl	2.6 M	–	–	en, tl, id, ms, vi
ALT [30]	18K	1,000	1,018	en, fil, id, ms, vi
IWSLT [6]	129K	1,553	1,268	en, vi

4 EXPERIMENTAL SETTINGS

We evaluated our extracted corpus on machine translation task, since we aim at utilizing this corpus for improving machine translation.

4.1 Datasets

We evaluated our *SeA-Wiki-Parl* corpus in machine translation task. For evaluation data, we employed the two existing manually translated corpora containing Southeast Asian languages, i.e., the Asian Language Treebank (*ALT*) corpus [30] and the *IWSLT* 2015 shared task data [6]. Both *ALT* and *IWSLT* are translated manually by native speakers, so these are quality datasets, and we employed them for our experiments. We present the data statistics in Table 6.[17]

4.1.1 SeA-Wiki-Parl. We used our extracted corpus containing more than 2.6M parallel sentences to train machine translation models for the Southeast Asian languges.

4.1.2 Asian Language Treebank (ALT). 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 20K English sentences and manually translated into the other languages by native speakers. We directly used the official splitting of training, development, and test sets provided by the authors.

4.1.3 IWSLT. This is a spoken language dataset containing subtitles in TED Talks, which are also manually translated and used for the well-known IWSLT machine translation shared task [6]. We used the IWSLT 2015 data[18] for English-Vietnamese translation. We used the official training set provided by the shared task. We used the *tst2012* for the development set, and *tst2013* for the test set, which are also officially provided by the shared task.

4.2 Task Settings

We setup two tasks based on the existing data sets, i.e., *ALT* task and *IWSLT* task.

(1) **ALT task**: we used the *ALT* development and test sets as presented in Table 6 for the ten language pairs between English, Filipino, Indonesian, Malay, and Vietnamese.
(2) **IWSLT task**: we used the *IWSLT* development and test sets as presented in Table 6 for English-Vietnamese and Vietnamese-English translations.

[17] In this paper, from this experiment part, we refer both Tagalog and Filipino as Filipino in evaluation, as we discussed the reason using Tagalog in Section 3.1.
[18] https://sites.google.com/site/iwsltevaluation2015/mt-track

4.3 Training Settings

In order to train translation models, we employed the *fairseq* model [27][19] implemented on Pytorch, which is a well-known Transformer-based model [43] and achieved state-of-the-art performance in various translation tasks [11, 27].

For training parameters, we used the provided basic architecture $transformer$ with "$-arch$" 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 [15], in which $adam - betas = (0.9, 0.98)$; $dropout = 0.1$; $weight\ decay = 0.0001$; and $max\ tokens = 4096$.

For evaluation, results are generated use beam search with $beam\ width = 5$ and $batch\ size = 128$. The scores are reported based on the *BLEU-4* from the *multi-bleu* available in the *fairseq*.

4.4 Compared Systems

We evaluated the performance of our extracted corpus by comparing the following translation systems, which mostly differ in terms of data settings.

(1) **Base-MT**: we use only the training set of the manual translated datasets, i.e., *ALT* or *IWSLT*, to train a translation model for each language pair. It is evaluated on its corresponding task, e.g., the Base-MT trained on the *ALT* training set is evaluated on the ALT task.
(2) **Wiki-MT**: we use only our extracted corpus to train a translation model for each language pair.
(3) **Enhanced-MT**: we merge the training set of the manual translated dataset with our extracted corpus to train a translation model for each language pair.

By comparing these models trained on different data settings, we would like to investigate the following questions.

(1) How good are the models trained on the manually translated data? (Base-MT)
(2) How good are the models trained on our automatically extracted corpus in comparison with the manually translated data? (Wiki-MT vs. Base-MT)
(3) How much our extracted corpus can contribute to existing corpora? (Enhanced-MT vs. Base-MT)

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

(1) **CVT** [10]: this is a semi-supervised learning algorithm improved from bidirectional LSTMs.
(2) **Stanford** [19]: this is a neural-based system participated in the IWSLT 2015 shared task, which is based on four LSTM-layer networks [13] and attention mechanism [20].

5 RESULTS AND ANALYSES

5.1 Results

We present the experimental results to answer the questions pointed out in Section 4.4. We first discuss the baseline results of Base-MT systems, which are trained on the manually translated training datasets. Then, we present the translation performance using our automatically extracted corpus. Finally, we show the results when enhancing the existing training data using our extracted corpus to investigate whether our corpus can provide new information.

5.1.1 Baseline results (Base-MT) on the Southeast Asian languages.

[19] https://github.com/pytorch/fairseq

Table 7. Results on the ALT task (test set, BLEU score; best scores are in bold, and scores are underlines when a Wiki-MT's score is higher than the Base-MT's score).

Source \ Target	System	English	Vietnamese	Indonesian	Malay	Filipino
English	Base-MT	–	20.37	25.57	30.57	17.77
	Wiki-MT	–	39.10	37.94	42.42	19.16
	Enhanced-MT	–	**39.82**	**40.18**	**45.61**	**28.74**
Vietnamese	Base-MT	11.47	–	7.84	9.42	**8.91**
	Wiki-MT	30.81	–	8.79	7.42	1.09
	Enhanced-MT	**33.29**	–	**17.27**	**18.12**	4.25
Indonesian	Base-MT	24.74	8.52	–	26.13	10.57
	Wiki-MT	35.02	7.71	–	25.34	0.86
	Enhanced-MT	**37.87**	**17.55**	–	**31.93**	**10.66**
Malay	Base-MT	27.02	10.36	27.57	–	**11.14**
	Wiki-MT	37.71	6.25	27.57	–	0.76
	Enhanced-MT	**41.11**	**15.92**	**32.71**	–	9.45
Filipino	Base-MT	22.10	**7.07**	**8.96**	**10.13**	–
	Wiki-MT	15.02	0.37	0.22	0.20	–
	Enhanced-MT	**28.58**	4.28	8.93	7.41	–

Table 8. Results on the IWSLT task (test set, BLEU score).

Model	English→Vietnamese	Vietnamese→English
Stanford [19]	26.9	–
CVT [10]	29.6	–
Base-MT	29.6	27.4
Wiki-MT	30.4	23.6
Enhanced-MT	**33.6**	**31.8**

ALT task. This research is the first investigation of MT task on these SeA languages such as the translations between Indonesian-Vietnamese, or Malay-Filipino, to our best knowledge. Therefore, based on the standard benchmark data such as the ALT corpus, we conduct experiments and report some first baseline results on these yet investigated languages. From the results of Base-MT systems presented in Table 7, we can observe several variations in performance, in which the scores are quite different between the language pairs. For instance, the results of English-Malay translations are very high (30.57, 27.02 BLEU), while the scores are much lower such as (Filipino-Vietnamese: 7.07 BLEU, Indonesian-Vietnamese: 8.52 BLEU, Vietnamese-English: 11.47 BLEU). It is noted that, all of these Base-MT systems are trained on the equivalent data (same training data size, extracted from the ALT corpus, and all of the non-English sentences are manually translated from the same original English sentences). Investigating these variations in terms of language aspects is not the main goal of this work, and we leave for future research.

IWSLT task. For the English-Vietnamese translations, this language pair has been evaluated in several previous work [6, 10, 19]. From the results reported in Table 8, our Base-MT obtained the comparable performance in comparison with CVT [10] and outperformed Stanford[19]. It is understandable when our MT systems are trained using the SOTA transformer-based fairseq model while Stanford [19] is trained on LSTM networks.

Table 9. Samples of extracted corpus (Non-English examples (*) are translated below). Samples (1, 2, 3, 4, 5) show correctly aligned sentences, while samples (5, 6, 7) show several incorrect alignment. The false aligned fragments are in red texts.

#	Languages	Source	Target
1	en-vi	It was released on 4 July 2014 by Monkey Puzzle and RCA Records worldwide, and Inertia Records in Australia.	Album được phát hành vào ngày 4 tháng 7 năm 2014 thông qua các hãng đĩa Monkey Puzzle và RCA Records trên toàn cầu và Inertia ở Úc.
2	en-id	The public DNS service and servers are maintained and owned by Cloudflare in partnership with APNIC.	Layanan DNS publik dan peladen dikelola dan dimiliki oleh Cloudflare yang bermitra dengan APNIC.
3	en-ms	The 1110s BC is a decade which lasted from 1119 BC to 1110 BC.	Dekad 1110-an SM merupakan tahun-tahun dari 1119 SM hingga 1110 SM.
4	id-ms(*)	Pada bulan April 2016, Aaron's, Inc. meraih penghargaan Guinness World Record untuk "permainan" domino kasur yang jatuh.	Pada April 2016, Aaron's, Inc. menetapkan Rekod Dunia Guinness untuk permainan domino tilam "terbesar".
5	ms-vi(*)	13 Reasons Why (makna harfiah: "13 Sebab Mengapa"; digayakan pada skrin sebagai TH1RTEEN R3ASONS WHY) adalah sebuah siri televisyen web drama remaja misteri Amerika Syarikat yang berdasarkan pada novel tahun 2007, "Thirteen Reasons Why" oleh Jay Asher dan diadaptasikan oleh Brian Yorkey untuk Netflix.	13 Reasons Why (cách điệu hóa trên phim thành Th1rteen R3asons Why) là một series truyền hình trực tuyến thuộc thể loại drama-thần bí của Mỹ dựa trên cuốn tiểu thuyết "Thirteen Reasons Why"năm 2007 của Jay Asher và được chuyển thể bởi Brian Yorkey cho Netflix.
6	tl-vi(*)	Ang 404 ay isang taon sa kalendaryong Gregorian. Ang 404 BC ay isang taon sa kalendaryong Gregorian.	Năm 404 là một năm trong lịch Julius. 404 TCN là một năm trong lịch La Mã.
7	id-vi(*)	13650 Perimedes (1996 TN49) adalah sebuah asteroid.	13650 Perimedes (provisional designation: 1996 TN) là một Trojan của Sao Mộc Hành tinh vi hình.
8	en-tl	Year 1002 (MII) was a common year starting on Thursday (link will display the full calendar) of the Julian calendar.	Ang 1002 ay isang taon sa kalendaryo.
9	en-vi	"Big Girls Cry" was released in June 2014.	Bài hát thứ hai, "Big Girls Cry", ra mắt vào tháng 6 năm 2014.

- (#4, id-ms): In April 2016, Aaron's, Inc. set a Guinness World Record for the "largest" domino game.
- (#5, ms-vi): 13 Reasons Why is an American teen drama television series developed for Netflix by Brian Yorkey and based on the 2007 novel Thirteen Reasons Why by author Jay Asher.
- (#6, tl-vi): 404 BC is a year in the Gregorian calendar.
- (#7, id-vi): 13650 Perimedes (1996 TN49) is an asteroid.

5.1.2 Results based on our automatically extracted corpus (Wiki-MT). We investigate the performance when using only our extracted corpus to train MT systems. Interestingly, from the results in Table 7, we found that Wiki-MT can even outperform Base-MT in several language pairs (8/20 cases) especially when the corpus size is large enough (usually

between a language paired with English). It is noted that Base-MT is trained on manually translated corpus while Wiki-MT is trained on our automatically annotated corpus.

ALT task. Specifically, all Wiki-MT systems containing English outperform Base-MT (except for Filipino-English). We achieve a significant improvement such as English-Vietnamese and Vietnamese-English (+18 BLEU), English-Indonesian and English-Malay (+12 BLEU) etc. However, the performance is low for Indonesian-Vietnamese (7.71), Malay-Vietnamese (6.25). Meanwhile, the results of most cases of Filipino are worst such as (Vietnamese-Filipino 1.09, Malay-Filipino 0.76). We found that there is a strong correlation of this performance with the size of our extracted corpus. As we presented the corpus size in Table 5), the number of parallel sentences containing English are from 300K to 800K, while those of Filipino are only less than 50K.

IWSLT task. For IWSLT results in Table 8, although the performance of Base-MT is reasonably high (29.6 BLEU), we still achieve an improvement (+0.8 BLEU) when training the MT system on only our automatically extracted corpus.

5.1.3 Results of enhancement by our extracted corpus. We have discussed the performance of MT systems trained on the existing manually annotated corpus and trained on only the extracted corpus. We would like to enhance the existing manually translated corpus by our extracted corpus (combining the two corpora for training MT systems). We investigate the two questions for the enhancement by comparing Base-MT and Enhanced-MT in Tables 7 and 8: 1) For the Base-MT systems when the results are already high, whether our extracted corpus can still help? 2) For the Wiki-MT systems with low performance, whether such cases are still helpful for the enhancement?

From the high-performance Base-MT systems. For the high performance Base-MT systems such as English-Malay (30.57), Malay-English (27.02), Vietnamese-English (27.4), or Indonesian-Malay (26.13), we can still obtain an improvement for the Enhanced-MT such as English-Malay (+15 BLEU), Malay-English (+14 BLEU), Vietnamese-English (+4 BLEU), or Indonesian-Malay (+5 BLEU). The results shows that our extracted corpus can still further improve the existing corpus, or in other words it provides helpful and new information.

From the lower-performance Wiki-MT systems. As we discussed in Section 5.1.2, the results of several Wiki-MT systems are lower than those of Base-MT such as Indonesian-Vietnamese (7.71 BLEU), Malay-Vietnamese (6.25 BLEU), or Filipino-English (15.02). However, we can still obtain the improvement with the Enhanced-MT on these language pairs, which indicate that our extracted corpus is still helpful to enhance the performance on the existing corpus.

Limited data. For the small size corpus of Filipino (less than 50K parallel sentences), the performance of Wiki-MT is very low. Therefore, when we combine the manually translated corpus with our extracted corpus to train Enhanced-MT systems, the performance decreased. This indicates that the extracted data containing Filipino is less reliable and harms the performance of the existing data. We will discuss this limitation in Section 5.2.4.

5.2 Analyses

5.2.1 Aligned Sentences. We analyze several issues from the extracted corpus, and present samples of the corpus in Table 9. The sentences are correctly aligned as presented in the samples (1, 2, 3, 4, 5, 6). For the sample (6) of Tagalog-Vietnamese, it is interesting to see *"Gregorian"* (in Filipino) is paired with different Vietnamese words "Julius" and "La Mã". These correctly aligned sentences make the corpus helpful, especially for improving MT systems.

Samples (7, 8, 9) show several problems of the aligned sentences. In samples (7) and (8), most words of a sentence is not correctly aligned with its corresponding sentence. In sample (9), the sentences are correctly aligned for most part of

Table 10. Analysis on vocabulary coverage. We present the vocabulary size and calculate the out-of-vocabulary (*OOV*) ratio of the *ALT*, *Wiki* (SeA-Wiki-Parl), and *ALT+Wiki* (merged ALT and SeA-Wiki-Parl).

#	Languages	Source					Target				
		#Vocabulary		OOV(%)			#Vocabulary		OOV(%)		
		ALT	Wiki	ALT	Wiki	ALT+Wiki	ALT	Wiki	ALT	Wiki	ALT+Wiki
1	English-Vietnamese	32K	466K	28.47	17.58	17.23	17K	404K	35.92	26.99	26.23
2	English-Indonesian	32K	262K	28.47	18.70	18.20	30K	266K	30.95	20.35	19.56
3	English-Malay	32K	341K	28.47	18.83	18.38	28K	341K	31.63	20.33	19.74
4	Indonesian-Malay	30K	100K	30.95	27.48	24.43	28K	101K	31.63	27.05	24.76
5	Indonesian-Vietnamese	30K	162K	30.95	26.85	23.82	17K	135K	35.92	31.95	30.26
6	Malay-Vietnamese	28K	153K	31.63	28.75	25.69	17K	138K	35.92	34.26	31.95
7	English-Filipino	32K	100K	28.47	25.24	22.15	41K	121K	35.00	32.03	26.38
8	Indonesian-Filipino	30K	45K	30.95	43.23	28.43	41K	50K	35.00	51.52	32.00
9	Malay-Filipino	28K	34K	31.63	46.98	29.68	41K	38K	35.00	58.13	32.56
10	Filipino-Vietnamese	41K	47K	35.00	56.71	32.35	17K	38K	35.92	41.22	34.14

the sentences, but a fragment of the Vietnamese sentence is redundant. For future work, we can improve the alignment performance by only extracting the correctly aligned fragments of the sentences, or filter out such wrong sentence pairs.

5.2.2 Vocabulary Coverage. One of the main factors affecting the MT performance is the vocabulary used to train the MT systems. In order to investigate the improvement when using our corpus to improve the Base-MT systems as presented in Table 7, we analyze the vocabulary coverage of the ALT, SeA-Wiki-Parl, and merged ALT with SeA-Wiki-Parl corpora, which are used to the train the Base-MT, Wiki-MT and Base+Wiki-MT respectively.

Out-Of-Vocabulary (OOV). We calculate the OOV ratio, which is the ratio of vocabulary in the evaluation set but not included in the training set. Higher OOV ratio means lower vocabulary coverage by the training data, which may decrease the MT performance [8, 21]. From the analyses presented in Table 10, we found that our extracted corpus can provide larger vocabulary set, which significantly reduces the OOV ratio from 1-10%. Comparing the OOV ratios of Wiki and ALT+Wiki, we observe that the difference is small in several cases (English-Vietnamese, English-Malay, Indonesian-Vietnamese), which indicates that using only our corpus can also significantly reduce the OOV ratio (without combining with the ALT corpus) in such cases. These analyses show that our corpus can provide larger and new vocabulary, which helps to reduce the OOV ratio, and leads to improve the MT performance.

5.2.3 Translation Output. We analyze several translation samples obtained from the models' translation output in Table 11 to further investigate the contribution of our corpus in terms of qualitative improvement.

For the first sample of Indonesian-Vietnamese, the Enhanced-MT generates a better translation *"flights were affected by the eruption"* while the Base-MT's output is *"flights were affected by the event"* which is less informative than the former.

For the second sample of English-Malay, the Wiki-MT also produces a correct translation *"four of the children"* while the Base-MT generates the output *"four of the trucks"* which is incorrect from the reference.

For the third sample of Vietnamese-English, the Enhanced-MT correctly translates "bài thuyết trình"into *"presentation"* when the Base-MT cannot translate this phrase, which is important to understand the meaning of this sentence.

Table 11. Translation samples (the overlapped (correctly translated) words and phrases between the predictions and the references are in underline)

Languages	Model	Sample	Meaning (in 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-MT (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 were affected by the event.
	Enhanced-MT(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-MT (ALT)	empat dari trak tersebut disuntik dari trak tersebut dan meninggal dunia di tempat kejadian.	four of the trucks were injected from the truck and died at the scene.
	Wiki-MT	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-MT (IWSLT)	what are the most powerful hand you can clapping in this morning?	
	Enhanced-MT (IWSLT)	what is the most clapping presentation in the morning?	

5.2.4 Limitations and Future Work. We have presented promising results of our corpus, and achieved the improvement in MT task on the SeA low-resource languages. The corpus provides larger and new vocabulary, which helps to train MT systems that are comparable or even outperform MT systems trained on manually translated corpora. However, although such promising results in MT task, there are still limitations remaining, which need to be further investigated and tackled in future work. We point out here several limitations as well as future directions below.

Improving and filtering aligned sentences. Since our corpus is extracted from available Wikipedia resources, parallel articles that we extracted can be seen as comparable data. In this kind of comparable data, each article pair may mention to the same content. However, regarding sentence unit, sentences in an article (of a language) may be only partially aligned with those in its aligned article (of the other language). Therefore, some sentence pairs in our corpus may not be entirely aligned such as the sample (9) in Table 9. Additionally, since our method is to automatically building corpora to overcome the data-hungry problem of the low-resource languages as well as time-consuming issue of manually constructing parallel data, it is difficult to avoid containing noise sentence pairs, which are mostly incorrectly aligned such as the sample (7) and (8) in in Table 9. Solving these two problems will improve the quality of our corpus. For future work, we plan to filter such noise sentence pairs by focusing on improving the sentence alignment phase. For the partial alignment, we intend to extract only the aligned fragments of the sentences rather than the entire sentence.

Further evaluating the corpus. Our main goal in this work is to build a corpus to improve MT task for the SeA languages. For this purpose, we have achieved the promising results and improvement on the ALT and IWSLT benchmark datasets. However, it is more interesting and helpful if we can filter noise sentence pairs and keep only high quality aligned sentences, and then it can be applied for other bilingual tasks rather than only MT. In order to do that, the corpus needs to be filtered and improved. In addition to automatically evaluating as we conducted in MT task in this work, we need to manually evaluate the corpus. We plan to conduct human evaluation in future work.

Limited Wikipedia resources. As presented in Table 7, the performance of MT systems containing Filipino languages are low, which are strongly related to the small corpus size (Table 5). The reason comes from the small data of the available Wikipedia resources for this language (Table 1). This limitation can be partly solved when the amount of Wikipedia articles is increasing.

Extending the corpus. Another direction is to extend the corpus to other languages and other bilingual tasks. In the scope of this work, we focus on the SeA languages to improve MT task. Since our method is automatically building parallel corpus from Wikipedia resources, it is possible and easy to extend the corpus to other languages. The improvement for MT task has been confirmed in this work. However, if the quality of this corpus can be improved, and human evaluation is conducted, we can extend the corpus to other bilingual tasks in these low-resource SeA languages. Providing such data resources will advance the development of natural language processing in general in these languages in the future.

6 CONCLUSION

In this work, we introduce a multilingual parallel corpus to improve machine translation for several Southeast Asian languages, which is automatically annotated from Wikipedia resources. The corpus contains more than 2.6 million parallel sentences ranging in ten language pairs Indonesian, Malay, Vietnamese, and Filipino, 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 limits the development of MT task on these languages. In order to build the corpus, we utilize the abundant available Wikipedia dumps resources, and the method includes three main steps. First, based on the available interlanguage link records from Wikipedia, we extract parallel titles, which are the titles of Wikipedia articles referring to the same content in different languages. Then, we extract the articles' texts based on these parallel titles. Finally, sentences in parallel articles are aligned to extract parallel sentences. In order to evaluate the contribution of the corpus to MT task, we conducted experiments on the two benchmark datasets, i.e., the Asian Language Treebank (ALT) and

the IWSLT shared task data. We achieved the improvement by using our corpus in both datasets, which confirms the contribution of our corpus in improving MT systems in the SeA languages. Additionally, we also conducted analyses to investigate the MT improvement as well as the issues of our corpus. The corpus can be improved and extended in several directions such as: filtering noise sentences, extracting high quality aligned fragments in sentence pairs, conducting human evaluation, extending to other bilingual tasks, or extending to other languages in future work.

REFERENCES

[1] Sisay Fissaha Adafre and Maarten De Rijke. 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*.

[2] Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A factory of comparable corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*. 3–13.

[3] Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovickỳ, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered. In *International Conference on Text, Speech, and Dialogue*. Springer, 231–238.

[4] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19, 2 (1993), 263–311.

[5] Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)* (28-30). Trento, Italy, 261–268.

[6] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*. Da Nang, Vietnam, 2–14. https://aclanthology.org/2015.iwslt-evaluation.1

[7] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[8] Chenhui Chu and Sadao Kurohashi. 2016. Paraphrasing out-of-vocabulary words with word embeddings and semantic lexicons for low resource statistical machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 644–648.

[9] 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. https://doi.org/10.1145/2833089

[10] Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-Supervised Sequence Modeling with Cross-View Training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1914–1925.

[11] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 489–500.

[12] Rajdeep Gupta, Santanu Pal, and Sivaji Bandyopadhyay. 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*. 69–76.

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Max Kaufmann. 2012. JMaxAlign: A maximum entropy parallel sentence alignment tool. In *Proceedings of COLING 2012: Demonstration Papers*. 277–288.

[15] Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR2015)*.

[16] 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

[17] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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, 48–54.

[18] Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. 104.

[19] Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*. 76–79.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1412–1421.

[21] Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 11–19.

[22] Mehdi Mohammadi and Nasser GhasemAghaee. 2010. Building bilingual parallel corpora based on wikipedia. In *2010 Second International Conference on Computer Engineering and Applications*, Vol. 2. IEEE, 264–268.

[23] Robert C Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*. Springer, 135–144.

[24] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 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)*. 314–319.

[25] Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. 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)*. 1–9.

[26] P. Gamallo Otero and Isaac Gonzalez Lopez. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*. 21–25.

[27] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 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)*. 48–53.

[28] Magdalena Plamadă and Martin Volk. 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, 112–120. https://www.aclweb.org/anthology/W13-2514

[29] 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

[30] Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, 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, 1–6.

[31] Motaz Saad, David Langlois, and Kamel Smaïli. 2013. Extracting comparable articles from wikipedia and measuring their comparabilities. *Procedia-Social and Behavioral Sciences* 95 (2013), 40–47.

[32] Rico Sennrich and Martin Volk. 2010. MT-based Sentence Alignment for OCR-generated Parallel Texts. In *Proceedings of AMTA 2010*.

[33] Sandhya Singh, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 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)*. 68–74.

[34] Jason R Smith, Chris Quirk, and Kristina Toutanova. 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, 403–411.

[35] 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.

[36] R Steinberger, B Pouliquen, A Widiger, C Ignat, T Erjavec, D Tufiş, and D Varga. 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)*.

[37] Liling Tan. 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, 184–193.

[38] Liling Tan and Francis Bond. 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*. 362–371.

[39] Liang Tian, Derek F Wong, Lidia S Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. 2014. UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation.. In *LREC*. 1837–1842.

[40] Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS.. In *Lrec*, Vol. 2012. 2214–2218.

[41] Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. *Proceedings of MT summit XI* (2007), 475–482.

[42] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*. 590–596.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[44] HV Vinayak, Fraser Thompson, and Oliver Tonby. 2014. Understanding ASEAN: Seven things you need to know. *growth* 2000 (2014), 13.

[45] Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics* 42, 2 (2016), 277–306.

[46] Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, Cheng Xiang Zhai, and Tie Yan Liu. 2019. Multi-agent dual learning. In *7th International Conference on Learning Representations, ICLR 2019*.

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

[48] Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 3530–3534.