Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5494–5502
Marseille, 20-25 June 2022

KC4MT: A High-Quality Corpus for Multilingual Machine Translation

Van-Vinh Nguyen, Ha Nguyen-Tien, Huong Le-Thanh, Phuong-Thai Nguyen, Van-Tan Bui
Nghia-Luan Pham, Tuan-Anh Phan, Minh-Cong Nguyen Hoang, Hong-Viet Tran, Huu-Anh Tran
Multilingual Machine Translation Project KC4.0, VNU - UET, Hanoi, Vietnam
vinhnv@vnu.edu.vn, tienhapt@gmail.com, bvtan@uneti.edu.vn, huonglt@soict.hust.edu.vn,
thainp@vnu.edu.vn, luanpn@dhhp.edu.vn, phantuananhkt2204k60@gmail.com,
congnhm@vnu.edu.vn, thviet79@gmail.com, anhuni1006@gmail.com

Abstract

The multilingual parallel corpus is an important resource for many applications of natural language processing (NLP). For machine translation, the size and quality of the training corpus mainly affects the quality of the translation models. In this work, we present the method for building high-quality multilingual parallel corpus in the news domain and for some low-resource languages, including Vietnamese, Laos, and Khmer, to improve the quality of multilingual machine translation in these areas. We also publicized this one that includes 500.000 Vietnamese-Chinese bilingual sentence pairs; 150.000 Vietnamese-Laos bilingual sentence pairs, and 150.000 Vietnamese-Khmer bilingual sentence pairs.

Keywords: Multilingual parallel corpus,low-resource languages, language resource, parallel corpus, machine translation

1. Introduction

In recent years, with the deep learning network, multilingual machine translation models have made leaps and bounds, the quality of the machine translation is close to that of a translator in some language pairs, such as French - English. Chinese-English. However, state-of-the-art machine translation models require high quality and large-scale parallel corpora for training to be able to reach near human-level translation quality (Wu and et al., 2016).

In machine translation, Vietnamese is known as a low-resource language. The public parallel corpora for Vietnamese are mainly English-Vietnamese bilingual corpus. Recently, A High-Quality and Large-Scale English-Vietnamese bilingual corpus was published by Vingroup, that called PhoMT (Doan et al., 2021). It includes 3.02M sentence pairs, it is very small, or even nonexistent in some language pairs, such as Vietnamese-Laos, Vietnamese-Khmer, etc.

For Vietnamese, the multilingual parallel corpus is very rare, (Trieu and Ittoo, 2020) introduced a multilingual parallel corpus containing 2.5 millions parallel sentences on ten language pairs of several Southeast Asian languages among Filipino, Malay, Indonesian, Vietnamese. However, this corpus is not publicly available for the research community.

China, Lao, Cambodia are countries bordering Vietnam. China is Vietnam's top trading partner, while Lao, Cambodia are two countries that have a friendly relationship that is considered by the Vietnamese government as brotherly love. A good quality machine translation system that translates text from Vietnamese into Chinese, Laos, or Cambodia is more essential than ever in order to support the information exchange of social-political organizations, economic groups, and people between countries. This has motivated us to focus on building a high-quality multilingual corpus, including Vietnamese, Chinese, Laos, and Cambodian languages to improve the quality of machine translation of these language pairs and publish available for the research community

The rest of the paper is laid out as follows: Section 2 presents the related works; Section 3 presents how we build multilingual parallel corpus; Section 4 presents experiment results, and we conclude and present future work in Section 5.

2. Related works

The multilingual parallel corpus is an electronic collection of texts in two or more languages put together in a principled way for the purpose of comparative linguistic studies and prepared in electronic form for search and analysis by computer (Dash and Selvaraj, 2018). It is an essential language resource for many applications of natural language processing, including multilingual machine translation. There has been much research on building and extending these ones.

(Tiedemann, 2016) built an OPUS that is a freely available sentence-aligned parallel corpora. OPUS covers over 200 languages and language variants with a total of about 3.2 billion sentences. It is collected from various sources and domains. Each sub-corpus in it is provided in common data formats to make it easy to integrate them in research and development.

The Oslo Multilingual Corpus (OMC) is a product of the interdisciplinary research project Languages in Contrast (SPRIK), which is a collaboration between researchers at the Faculty of Humanities, University of Oslo.[1]. It is an extension of the ENPC that is a bidirectional translation corpus consisting of original En-

[1]https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/

glish texts and their translations into Norwegian, and Norwegian original texts and their translations into English was built in the 1990s. The OMC contains many sub-corpora that differ in composition with regard to languages and number of texts included. It is mainly the languages Norwegian, English, French, and German that are represented in the sub-corpora, but some of the corpora include Dutch and Portuguese texts. In addition, there are related parallel corpora for English-Swedish and English-Finnish, compiled in Sweden and Finland, which are accessible from the same site.

(Salesky et al., 2021) released the Multilingual TEDx corpus to facilitate speech recognition and speech translation research. This corpus is a collection of audio recordings from TEDx talks in 8 source languages. They segment transcripts into sentences and align them to the source language audio and target-language translations. It is built to support speech recognition and speech translation research across many non-English source languages.

(Soares and Krallinger, 2019) presented the development of parallel corpora from BVS (Health Virtual Library) in three languages: English, Portuguese, and Spanish. Sentences were automatically aligned using the Hunalign algorithm for EN/ES and EN/PT language pairs, and for a subset of trilingual articles also. They demonstrate the capabilities of their corpus by training a Neural Machine Translation system for each language pair, which outperformed related works on scientific biomedical articles. The sentence alignment method is used in this work gets an average of 96% correctly aligned sentences across all languages. Their parallel corpus is freely available, with complementary information regarding article metadata.

(Siripragada et al., 2020) presented the methods of constructing sentence-aligned multilingual parallel corpora using tools enabled by recent advances in machine translation and cross-lingual retrieval using deep neural network based methods. This corpora includes 10 Indian languages: Hindi, Telugu, Tamil, Malayalam, Gujarati, Urdu, Bengali, Oriya, Marathi, Punjabi, and English. Among them, there are some languages which are known to be low-resource languages. It is collected from online sources that have content shared across languages. These corpora were built in the context that some languages are either not large enough or restricted to a specific domain.

Our method of constructing sentence-aligned multilingual parallel corpora differs from previous works as follows:

- How to organize text pairs crawled from online sources in order to find the most parallel text pairs;
- We aim to build high-quality multilingual parallel corpora on News domain;
- Our corpora includes four languages: Chinese, Vietnamese, Laos, Khmer. But Vietnamese, Laos, Khmer are low-resources.

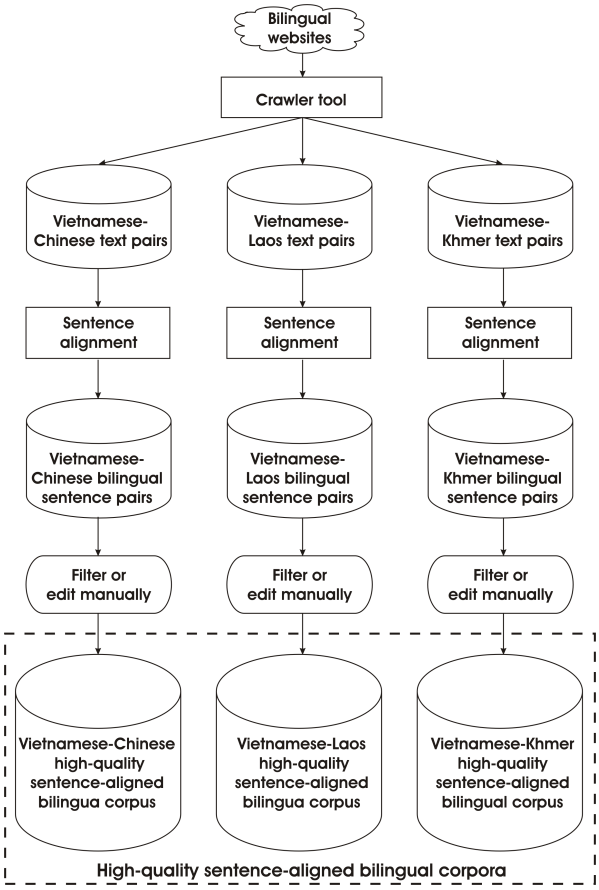

Figure 1: The method of constructing high-quality sentence-aligned multilingual parallel corpus.

In next section, we are going to present our method in constructing this corpora.

3. Building Multilingual parallel Corpus

Pages in html type are downloaded from bilingual websites using a crawler tool. These pages are pre-processed to get parallel text pairs. Then we do text alignment, paragraph alignment, and finally, sentence alignment by using tools we designed ourselves to get parallel sentence pairs. Finally, we review manually these pairs to get high quality parallel corpora. Our proposed method for constructing high-quality parallel corpora is shown in Figure 1

3.1. Collecting Parallel Text Pairs

We crawl bilingual websites in the news domain for Vietnamese-Chinese, Vietnamese-Laos, and Vietnamese-Khmer language pairs. It includes: ”tapchicongsan.org.vn”, ”vietnamplus.vn” ”tapchilaoviet.org”,”nhandan.vn”, ”dantocmiennui.vn”, ”dangcongsan.vn”, ”cantho.gov.vn”, ”travinh.gov.vn”, ”baotravinh.vn”, ”tvu.edu.vn”. We extract only the text from html. Depending on posted date and categories of these html pages in order to organize and store the collected text pairs on both the source and destination sides.

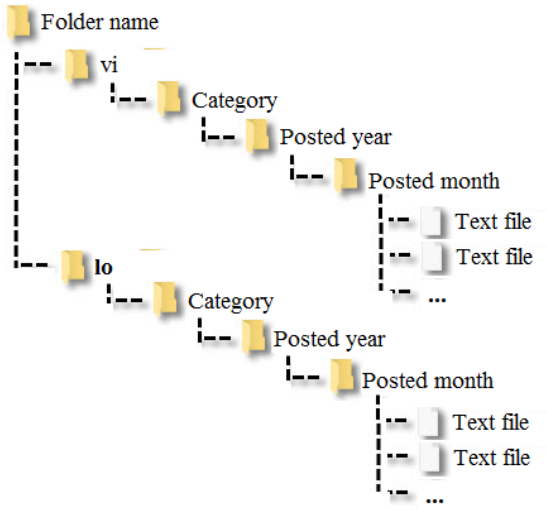

Figure 2: The text stored tree for Vietnamese - Laos pair

Figure 2 shows the tree that stores text crawled from Vietnamese - Laos websites. Similar trees were created for the Vietnamese-Chinese and Vietnamese-Khmer language pairs.

Many news texts are not translated into another language on the same day, but they are mostly posted in the same month instead. Therefore, when we crawl them from bilingual websites, they are stored by posted month. We only compare texts within the same month for each language pair, in order to reduce the complexity but not lose potential candidate document pairs.

3.2. Sentence Alignment

The sentence alignment task takes parallel text pairs as the input and returns bilingual sentence pairs. It can be done by creating all candidate sentence pairs from bilingual documents and then computing the semantic similarity between these sentence pairs. In the past, some researchers translated both documents to a third language that is a rich-resources and popular one, such as English, then used a similarity score such as Levenstein or cosine to compute their similarity (Abdul-Rauf and Schwenk, 2009) (Sennrich and Volk, 2011). Recent studies use a multilingual model, such as BERT or LASER, to represent sentence embedding vectors and compute cosine similarity between these vectors (Grégoire and Langlais, 2018) (Artetxe and Schwenk, 2019) (Thompson and Koehn, 2019) (Chousa et al., 2020). Although these methods have shown superior performance compared to previous ones, it requires large training data. This is the biggest difficulty when applying them for low-resource language pairs. In addition, the learned models sometimes do not cover all the language's features.

Besides the above weakness, the state-of-the-art method Vecalign also has other limitations. First, it cannot align pairs of sentences that are located far apart in the source and target documents. Second, two sentences that are not translations of each other but have a highly similarity still be aligned by Vecalign. In this work, we propose a sentence alignment method that overcomes the above-mentioned limitations by taking advantages of machine translation tools and improving Vecalign.

Our proposed method as follows: First, we do document alignment based on the similarity of two document embedding vectors and the length ratio of the document pair. Then, we do paragraph alignment for each aligned text pair based on the similarity of two paragraph embedding vectors and the length ratio of the paragraph pair. Finally, for each aligned paragraph pair, we segment them into sentences, then do sentence alignment based on the similarity of the two sentence embedding vectors and length ratio of the sentence pair. In our above proposed method, we use the length ratio of two spans (i.e., text, paragraph, sentence) in alignments to overcome Vecalign's limitation when text span pairs have high Cosine similarity of embedding vector pair but not in alignment. An example of such a pair of Vietnamese-Laotian sentences is shown in Figure 3. And we do paragraph alignment before sentence alignment to limit the spread of alignment errors from one paragraph to another, and to overcome Vecalign's limitation when faced with sentence alignment, that are translations of each other are located far apart in the text. Besides, we use machine translation tools to convert low-resource languages into rich-resource languages that can apply deep learning models for them.

In this work, we use the publicly available LASER multilingual sentence embedding method (Artetxe and Schwenk, 2018) and model, which is pre-trained on 93 languages[2]. For languages that aren't in these 93 languages, we use deep-translator[3] to translate it into one of 93 languages and use LASER after.

The similarity of two text spans is computed based on the cosine similarity (Garcia, 2015) between their embedding vectors.

The length ratio of two text spans is computed as the number of characters in one text span divided by the number of characters in the other one. Text, paragraph, sentence embedding vectors v_t, v_p, v_s are computed as follows:

- v_s is built based on LASER.
- $v_p = \frac{\sum_{i=1}^{n} v_s[i]}{n}$

  Where: n is the number of sentences in paragraph.
- $v_t = \frac{\sum_{j=1}^{k} v_p[j]}{k}$

  Where: k is the number of paragraphs in text.

[2]https://github.com/facebookresearch/LASER
[3]https://github.com/nidhaloff/deep-translator

Vecalign's similarity	Vietnamese sentence	Laos sentence
0.81	Merchant muốn tăng trưởng doanh thu bền vững thì cần thiết phải thực sự hiểu khách hàng của mình (*Merchant want sustainable revenue growth, they need to really understand their customers*)	ຮ້ານຄ້າຕ້ອງການເພີ່ມລາຍໄດ້ຢ່າງຍືນຍົງນັ້ນ ຈຳເປັນຕ້ອງເຂົ້າໃຈລູກຄ້າຂອງຕົນເອງຢ່າງເລິກເຊິ່ງ, ຕ້ອງຮູ້ລັກສະນະພິເສດຂອງລູກຄ້າ ແລະ ຕ້ອງຮູ້ເຖິງພຶດຕິກຳການບໍລິໂພກຂອງພວກເຂົາ ຈາກຂໍ້ມູນເຫຼົ່ານັ້ນ ແມ່ນສາມາດນຳມາໃຊ້ໃນການປັບປຸງຜະລິດຕະພັນຂອງຄ້າ, ລາຄາ, ວິທີການບໍລິການ ຫຼື ການສ້າງໂປຣໂມຊັ່ນກະຕຸ້ນຍອດຂາຍໃຫ້ເໝາະສົມກັບລູກຄ້າໃຫ້ຫຼາຍທີ່ສຸດ. (*Stores want sustainable revenue growth, they need to really understand their customers, it is about the private characteristics of customers and their consumption behavior.*)

Figure 3: Pair of sentences are aligned by vecalign

Where: $v_p[i]$ is i^{th} paragraph embedding vector.

The text and paragraph alignment: The text and paragraph alignments are performed by using a brute-force algorithm, as described below.

- *Text alignment*: Two texts in two folders with the same month and the same categories in both languages are aligned if they satisfy the following conditions:
  1. The Cosine similarity of the text embedding vectors is greater than a threshold α.
  2. The length ratio of two texts is in (β, γ)
- *Paragraph alignment*: Two paragraphs in each aligned text pair are aligned if they satisfy the following conditions:
  1. The Cosine similarity of the paragraph embedding vectors is greater than a threshold θ.
  2. The length ratio of two paragraphs is in (δ, ε)

Where: $\alpha.\beta, \gamma, \theta.\delta, \varepsilon$ are found depending on each language pair.

The threshold values $\alpha.\theta$ depend on each language pair and are determined manually by experimenting on the sample data set for that language pair. α is chosen as 0.8 for Vietnamese-Laos, Vietnamese-Chinese language pairs, and 0.75 for the Vietnamese-Khmer language pair; θ is chosen as 0.75 for Vietnamese-Laos, Vietnamese-Chinese language pairs, and 0.7 for the Vietnamese-Khmer language pair.

β, $\gamma, \delta, \varepsilon$ are the minimum and maximum character-based length ratio between two documents and two paragraphs, respectively. They are estimated based on statistics on the given sentence-aligned bilingual corpus. β and δ are chosen as 0.7 for Vietnamese-Laos, Vietnamese-Khmer language pairs, and 0.3 for the Vietnamese-Chinese language pair; γ and ε are chosen as 1.3 for Vietnamese-Laos, Vietnamese-Khmer language pairs, and 0.8 for the Chinese-Vietnamese language pair.

Sentence alignment: *Vecalign*[4] is the improved sentence alignment method in Linear Time and Space (Thompson and Koehn, 2019). It does sentence alignment based on the similarity of bilingual sentence embeddings. In this work, we improved it by using a character-based length ratio between the source sentence and target sentence and adding a paragraph alignment phase before sentence alignment to prevent the propagation of sentence alignment errors from one paragraph to another. On average, our proposed sentence alignment method outperforms Vecalign by 8,64 %.

Our proposed sentence alignment method got 98% precision for Vietnamese - Laos, Vietnamese - Khmer and Vietnamese - Chinese pairs.

3.3. Manual Data Reviewing

We built the online tool[5] for manual data reviewing for automatic aligned bilingual sentences. Figure 4 showed the reviewing interface for the Vietnamese-Laos pair. Each bilingual sentence pair is designed with two reviewing levels, including:

- Good level: The bilingual sentence pair is selected at this level by annotators if they are exactly translation of each other. For bilingual sentence pairs that are easy to modify to get good pairs, annotators will select good level for it after modifying them.
- Bad level: The bilingual sentence pair is selected at this level by annotators if they are not translation of each other or difficult to modify to get good pairs.

If a bilingual sentence pair is evaluated as good, it will be added to the corpus. It will be removed otherwise. Our goal is to build a high-quality multilingual parallel corpus, so we have chosen good annotators for manual data reviewing and use the best expert to randomly review 10% of the manually reviewed data every month for fine quality control of the data added to the corpus. We recruited 35 annotators, including 15 for the Vietnamese-Chinese pair and 10 for each of the Vietnamese-Laos and Vietnamese-Cambodian, who manually reviewed bilingual sentence pairs from the automatically collected ones. These annotators speak and write fluently in the language pair that they are assigned to review. They were final-year undergraduate students, graduate students, or teachers at universities.

[4]https://github.com/thompsonb/vecalign

[5]http://nmtuet.ddns.net:3000

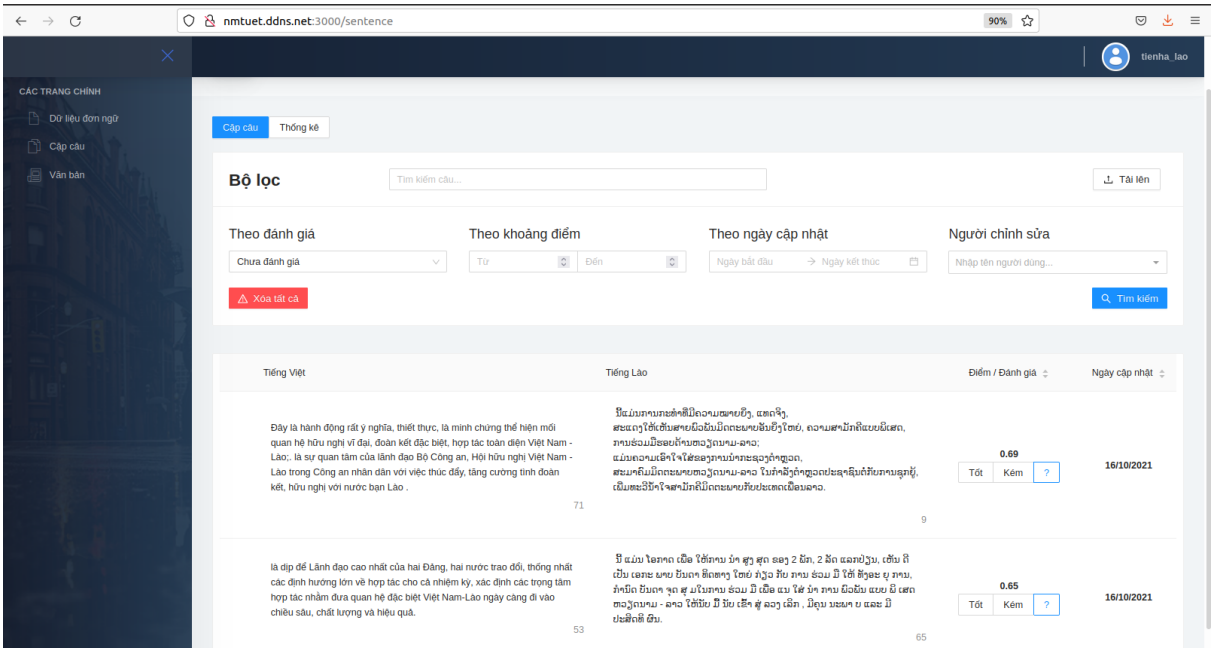

Figure 4: The online tool for manual Vietnamese-Laos data reviewing

Firstly, all annotators were trained to use an online tool for manually reviewing and knowing how to choose good or bad for each sentence pair. Each annotator was provided with an account to review daily. On average, each day, one annotator spent about two hours reviewing the data, and it took one year to complete this corpus.

4. Experiments

In this section, we carry out statistics and experiments to evaluate the quality of our multilingual parallel corpus. Specifically, section 4.1 presents statistical results on this one. Section 4.2 describes intrinsic evaluations that are the assessment of language experts on the quality of bilingual sentence pairs. Finally, Section 4.3, presents extrinsic evaluations using a multilingual neural network machine translation model.

4.1. Parallel Corpus Statistics

We carry out statistics on parallel corpora including vocabulary, number of bilingual sentence pairs, maximum and minimum length of sentences, average length of sentences in distinct languages, mean deviation in sentence length between two languages, illustrating the distribution of sentences according to the length of the distinct languages.

Our multilingual parallel corpus includes three language pairs: Vietnamese-Laos (Lo-Vi), Vietnamese-Khmer (Kh-Vi), Vietnamese-Chinese (Zh-Vi). The number of sentences and words for each one are shown in Table 1, where S is the number of sentences, $T1$ and $T2$, $V1$ and $V2$ are the token numbers, and the vocab size of the first and second language, respectively. It is publicized at "https://github.com/KCDichDaNgu/MultilingualMT-UET-KC4.0".

Table 1: Statistics of sentences, tokens, and vocabulary from our corpus.

Corpus	#S	#T1	#V1	#T2	#V2
Lo-Vi	150K	3,693K	60K	3,517K	61K
Kh-Vi	150K	4,329K	53K	4,189K	53K
Zh-Vi	500K	10,598K	70K	9,460K	73K

The statistical results for sentence length in terms of words in our corpus are shown in Table 2, in which Ma, Mi, Avg are the maximum, minimum, and average values of sentence length, respectively. Δ is the mean deviation between two bilingual sentences. *Language 2* is Laos, Khmer, or Chinese. Figure 5 illustrates the distribution of sentence length by language.

Table 2: Statistics on the length of sentences in our corpus.

Corpus	Vietnamese			Language 2			Δ
	Ma	Mi	Avg	Ma	Mi	Avg	
Lo-Vi	157	1	24.3	152	1	23.5	0.8
Kh-Vi	148	1	28.7	139	1	27.4	1.3
Zh-Vi	176	1	21.2	165	1	19.8	1.4

4.2. Intrinsic Evaluations

We randomly selected 1000 bilingual sentence pairs for each language pair from our corpus to evaluate manually by language experts. Each sentence pair is rated by three annotators on the semantic similarity by a value

Table 3: Annotation guidelines provided to annotators.

Title	Scale	Description
Very Good	4	Two sentences are completely similar in meaning. Two sentences that refer to the same object or concept, using words that have semantic similarity or synonyms to describe them. The length of the two sentences is equivalent (similarity score from 90 to 100).
Good	3	Two sentences with similarities in meaning, referring to the same object or concept. The length of the two sentences may vary slightly (similarity score from 70 to 89).
Need correction	2	Two sentences that are related in meaning, each referring to objects or concepts but they are related. The length of two sentences may vary slightly (similarity score from 50 to 69).
Bad	1	Two sentences that are different in meaning but have a slight semantic related, may share the same topic. The length of two sentences can vary greatly (similarity score from 30 to 49).
Very Bad	0	The two sentences are completely different in meaning, their content is not related to each other. The length of two sentences can vary greatly (similarity score from 0 to 29).

between 0 and 100. These evaluation results are used to evaluate the quality of this corpus, as well as the consensus among annotators through kappa and correlation coefficients. To evaluate the consensus among annotators accurately, we map the semantic similarity of sentence pairs to five quality levels including "*Very good*", "*Good*", "*Needs correction*", "*Bad*", and "*Very bad*". The evaluation results of the annotators can show the quality of our corpus in terms of internal evaluations.

4.2.1. Annotators

The set of bilingual sentence pairs in each language pair is assessed by three annotators. The annotators are provided guidelines as shown in Table 3, their assessment is carried out independently.

4.2.2. Intrinsic Evaluation Results

The results of the annotators' quality assessment of sentence pairs are presented in Table 4, in which #4, #3, #2, #1, and #0 correspond to levels *Very Good*, *Good*, *Needs correction*, *Bad*, and *Very Bad*, respectively. The *ASS* column of this table shows the average similarity score of sentence pairs that are rated by the annotators. The results in Table 4 are given as percentages. To estimate the reliability of the annotators' evaluation results, we use inter-annotator agreement scores, whereby Fleiss' kappa coefficient (Fleiss, 1971) and Spearman correlation are used. The Consensus evaluation results are presented in Table 5.

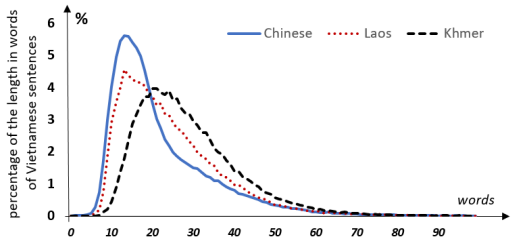

Figure 5: An illustration of statistical results on sentence length in parallel corpora.

Table 4: Synthesize the quality evaluating results of sentence pairs.

pair	#4	#3	#2	#1	#0	ASS
Lo-Vi	61.7	31.1	7.0	0.2	0.0	91.3
Kh-Vi	9.4	61.0	24.6	4.1	0.9	72.8
Zh-Vi	68.9	18.9	6.8	3.9	1.6	87.4

Table 5: Consensus Score of annotators.

Language pair	Kappa	Spearman
Lo-Vi	0.72	0.79
Kh-Vi	0.76	0.82
Zh-Vi	0.81	0.89

4.3. Extrinsic Evaluations

Our goal is to build a high-quality multilingual parallel corpus for Multilingual Machine Translation, so we use neural MT systems for extrinsic evaluations. We conduct experiments on our multilingual dataset to study: (i) a comparison between the well-known automatic translation engine (here, Google Translate) and neural MT baseline systems, and (ii) evaluate the quality of our multilingual dataset. So this section, we describe different experiment scenarios. We train multilingual NMT models, in which one model using our multilingual dataset, and another one using the ALT Parallel Corpus [6] for baseline *(this corpus under Asian Language Treebank (ALT) Project aims to advance the state-of-the-art Asian natural language processing techniques)*. Then we evaluate the quality

[6]https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/

of models on the ALT test set.

We also train bilingual NMT models for each language pair when varying training sizes, then we evaluate on the our test sets with respect to sentence lengths of reference Vietnamese sentences. We report standard metric BLEU (Papineni et al., 2002), this metric is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another.
The detailed statistics for the datasets are described in the Table 6 and in the Table 7.

Table 6: Our Bilingual datasets

Language pairs	Train	Valid	Test
Zh-Vi	500k	1000	2002
Lo-Vi	150k	1000	2002
Kh-Vi	150k	1000	2002

Table 7: The ALT Parallel Corpus datasets

Language pairs	Train	Valid	Test
Zh-Vi	18k	1000	1018
Lo-Vi	18k	1000	1018
Kh-Vi	18k	1000	1018

Preprocessing All parallel texts were tokenized and truncated using sentencepiece scripts, and then they are applied to Sennrich's BPE (Sennrich et al., 2016). We explore 32K operators are learned to generate BPE codes for all languages. For Vietnamese, we only use Moses's scripts for tokenization and true-casing.

Systems and Training We implement our NMT system [7] from a zeros-base to train all our experiments. The same settings are used for all experiments.
We trained our Transformer model [8] using the number of encoder 12, decoder layers are 6, 8 head is used, d_{model} is 512, dropout value is 0.1, batch size of 64, learning rate value is 0.4 with the aid of Adam optimizer. The learning rate has warmup updates by 8000 steps and label smoothing value is 0.1. We evaluate the quality of two systems *(1) Bilingual system*, *(2) Multilingual system*.

(1) *Bilingual system*.We train systems on our separate bilingual data and the ALT Parallel Corpus for each language pair. We utilized the best model to decode the test data for comparison purposes of our experiments. We train Chinese-Vietnamese, Laos-Vietnamese, and Khmer-Vietnamese models for 20 epochs. We train bilingual models when varying training sizes, and compared the quality of them with Google Translate on the test sets with respect to sentence lengths of reference Vietnamese sentences.

(2) *Multilingual system*. We concatenate bilingual datasets for all language pairs in order to construct the new datasets: Chinese, Khmer, Laos, Vietnamese. We get two new multilingual datasets, one is the multilingual ALT system and another one is our multilingual system. We train two multilingual systems on above multilingual datasets for the same number of epochs.

Results The experiment results are shown in the Table 8, 9 for bilingual systems, and the Table 10 for multilingual systems.

Table 8 presents BLEU scores obtained by the automatic translation engines and our neural MT system on the test sets with respect to sentence length bucket for the Khmer-to-Vietnamese (Kh-to-Vi), Laos-to-Vietnamese (Lo-to-Vi) and Chinese-to-Vietnamese (Zh-to-Vi) translation setups. In Table 8, we find that models produce higher BLEU scores for short- and medium-length sentences (i.e. $<$ 30 tokens) than for long sentences. This is not surprising as a major proportion of short and medium length sentences.

In Table 9, we compare bilingual systems including ALT system and our system. BLEU scores of our system are all higher than ALT systems. Here, our models obtain 10.3+ to 17.1+ points absolute better than ALT model, which corresponds to a relative performance improvement (Δ) from 86.2% to 201.3%.

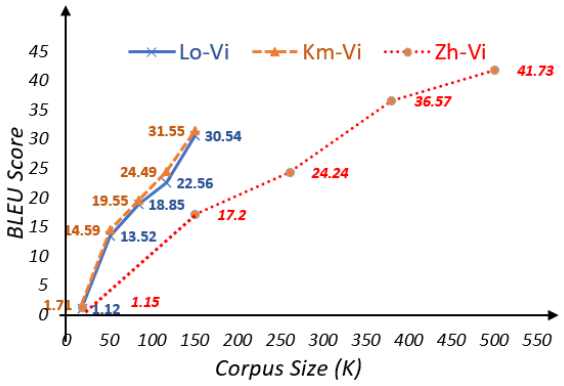

Figure 6: BLEU scores of our bilingual NMT system on the Zh-to-Vi, Lo-to-Vi, Kh-to-Vi test set when varying training sizes.

Figure 6 presents BLEU scores of our model on the test sets for the Zh-to-Vi, Lo-to-Vi, Kh-to-Vi setup when varying the numbers of training sentence pairs. Those scores clearly show the effectiveness of larger training sizes. Thus this experiment also reconfirms the positive effect of a larger training size.

In Table 10, we compare multiligual systems including ALT system and our system. BLEU score of our multilingual system is higher than multilingual ALT system on the same test set. Here, our models obtain 14.51+ to 16.27+ points absolute better than ALT model.

[7] https://kcdichdangu.ddns.net:3001/
[8] https://arxiv.org/pdf/2112.15272.pdf

Table 8: BLEU scores on the test set with respect sentence lengths of reference Vietnamese sentences.

Model		BLEU scores/Sentence length					
		(0, 10)	**[10, 20)**	**[20, 30)**	**[30, 40)**	**[40, 50)**	**[50, +inf)**
Kh-to-Vi	Sentence number	3	257	770	627	223	122
		(0.15%)	(12.84%)	(38.46%)	(31.32%)	(11.14%)	(6.09%)
	Our system	**0.00**	**30.40**	**31.37**	**30.54**	**28.18**	**19.59**
	Google Translate	0.00	51.74	58.73	57.19	54.33	44.64
Lo-to-Vi	Sentence number	0	427	739	530	186	120
		(0.00%)	(21.33%)	(36.91%)	(26.47%)	(9.29%)	(5.99%)
	Our system	**0.00**	**31.95**	**30.38**	**29.66**	**29.43**	**21.90**
	Google Translate	0.00	46.47	46.55	45.23	44.27	40.03
Zh-to-Vi	Sentence number	1	501	865	357	159	119
		(0.05%)	(25.02%)	(43.21%)	(17.83%)	(7.94%)	(5.94%)
	Our system	**0.00**	**40.55**	**40.75**	**38.60**	**38.28**	**37.35**
	Google Translate	0.00	41.98	44.91	45.04	46.98	44.99

Table 9: Overall results with respect to Bilingual systems. BLEU score of the system when trained with ALT corpus and our corpus, evaluated on the ALT testing dataset.

Pairs	**BLEU scores**		
	ALT system	**Our system**	$\Delta(\%)$
Zh-Vi	8.47	25.52	201.3
Lo-Vi	8.59	22.78	165.2
Kh-Vi	11.97	22.29	86.2

Table 10: Overall results with respect to multilingual machine translation systems. BLEU score of the system when trained with ALT corpus and our corpus, evaluated on the ALT testing dataset.

Language pairs	**BLEU scores**	
	ALT system	**Our system**
Zh-Vi	11.77	28.04
Lo-Vi	10.40	24.91
Kh-Vi	12.79	28.87

5. Conclusion and Future Work

In this paper, we have presented the method for building high-quality multilingual parallel corpora in the news domain and have shared it for free. Our corpora are great value for low-resource languages such as Vietnamese, Laos, and Khmer. We also deployed some experimentally to test the quality of these corpora. It improved by an average of 11.37 BLEU when added to the corpus for training neural machine translation systems. In the future, we will continue to expand this corpus in both size and number of language pairs. Furthermore, we will conduct research to use our corpus to improve the quality of multilingual machine translation systems and some applications of NLP.

Acknowledgments

This work has been supported by Ministry of Science and Technology of Vietnam under Program KC 4.0, No. KC-4.0.12/19-25.

6. Bibliographical References

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve smt performance. In *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 16–23, 01.

Artetxe, M. and Schwenk, H. (2018). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

Artetxe, M. and Schwenk, H. (2019). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Chousa, K., Nagata, M., and Nishino, M. (2020). SpanAlign: Sentence alignment method based on cross-language span prediction and ILP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Dash, N. and Selvaraj, A., (2018). *Limitations of Language Corpora*, pages 259–272. 01.

Doan, L., Nguyen, L. T., Tran, N. L., Hoang, T., and Nguyen, D. Q. (2021). Phomt: A high-quality and large-scale benchmark dataset for vietnamese-english machine translation.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Garcia, E. (2015). Cosine similarity tutorial. *Published: 04-10-2015; Updated: 09-15-2018 © Edel Garcia, PhD; admin@minerazzi.com*, 04.

Grégoire, F. and Langlais, P. (2018). Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In *Proceedings of the 27th International Conference on Com-*

putational Linguistics, pages 1442–1453, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D. W., and Post, M. (2021). The multilingual tedx corpus for speech recognition and translation. *CoRR*, abs/2102.01757.

Sennrich, R. and Volk, M. (2011). Iterative, mt-based sentence alignment of parallel texts. In *NODALIDA 2011, Nordic Conference of Computational Linguistics*. Northern European Association for Language Technology (NEALT), May. The 18th Nordic Conference of Computational Linguistics ; Conference date: 11-05-2011 Through 13-05-2011.

Sennrich, R., Haddow, B., and 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)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Siripragada, S., Philip, J., Namboodiri, V. P., and Jawahar, C. V. (2020). A multilingual parallel corpora collection effort for indian languages. *CoRR*, abs/2007.07691.

Soares, F. and Krallinger, M. (2019). BVS corpus: A multilingual parallel corpus of biomedical scientific texts. *CoRR*, abs/1905.01712.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, November. Association for Computational Linguistics.

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia, May 30–June 1. Baltic Journal of Modern Computing.

Trieu, H. and Ittoo, A. (2020). Mt-wiki: A wikipedia-based multilingual parallel corpus for machine translation on low-resource languages. volume Published 2020.

Wu, Y. and et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.