IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2022, accepted 8 June 2022, date of publication 29 June 2022, date of current version 1 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3186978

A Method of Chinese-Vietnamese Bilingual Corpus Construction for Machine Translation

PHUOC TRAN[1], THIEN NGUYEN[1], DINH-HONG VU[1], HUU-ANH TRAN[2], AND BAY VO[3]

[1]Natural Language Processing and Knowledge Discovery Laboratory, Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam
[2]Faculty of Information Technology, Thai Binh University, Thai Binh City 410000, Vietnam
[3]Faculty of Information Technology, HUTECH University, Ho Chi Minh City 700000, Vietnam

Corresponding author: Bay Vo (vd.bay@hutech.edu.vn)

ABSTRACT A bilingual corpus is vital for natural language processing problems, especially in machine translation. The larger and better quality the corpus is, the higher the efficiency of the resulting machine translation is. There are two popular approaches to building a bilingual corpus. The first is building one automatically based on resources that are available on the internet, typically bilingual websites. The second approach is to construct one manually. Automated construction methods are being used more frequently because they are less expensive and there are a growing number of bilingual websites to exploit. In this paper, we use automated collection methods for a bilingual website to create a bilingual Chinese-Vietnamese corpus. In particular, the bilingual website we use to collect the data is the website of a multilingual dictionary (https://glosbe.com). We collected the Chinese-Vietnamese corpus from this website that includes more than 400k sentence pairs. We chose 100,000 sentence pairs in this corpus for machine translation experiments. From the corpus, we built five datasets consisting of 20k, 40k, 60k, 80k, and 100k sentence pairs, respectively. In addition, we built five additional datasets, applying word segmentation on the sentences of the original datasets. The experimental results showed that: 1) the quality of the corpus is relatively good with the highest BLEU score of 19.8, although there are still some issues that need to be addressed in future works; 2) the larger the corpus is, the higher the machine translation quality is; and 3) the untokenized datasets help train better translation models than the tokenized datasets.

INDEX TERMS Construction of a bilingual corpus, Chinese-Vietnamese machine translation, dictionary websites, Glosbe.

I. INTRODUCTION

A bilingual corpus is a basic requirement for building a machine translation system, whether it is statistical machine translation, typically a phrase-based statistical machine translation (PSMT), or a neural machine translation (NMT). The larger and better quality the bilingual corpus is, the better the translation results are. There are currently large bilingual corpora available to the research community, especially for resource-rich language pairs such as English-Chinese [1], English-German [2], [3] and so on. However, for low-resource language pairs, like Chinese-Vietnamese, having a large and quality bilingual corpus is impossible at the moment. Therefore, the creation of a Chinese-Vietnamese bilingual corpus for the research community, particularly research into Chinese-Vietnamese machine translation, is extremely important.

There are two common approaches to building a bilingual corpus, which are: (1) building by automatically collecting a bilingual corpus on the internet [4], and (2) building a bilingual corpus by manually collection [5]. For (1), the system needs access to bilingual websites, as it automatically extracts bilingual sentence pairs from bilingual articles. This method requires the fully parallel documents that exist on bilingual websites. For (2), to get a bilingual corpus we need bilingual experts to input and edit data. Obviously with this manual method we need to spend a great deal of money to hire language experts to carry out the work, although the quality

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

of the resulting bilingual corpus will be much better than with the automated method.

Currently, to the best of our knowledge, there are about eight Vietnamese-Chinese bilingual websites, including Thoi Dai's website (http://thoidai.com.vn/), Nhan Dan's website (http: //www.nhandan.com.vn/), VietnamPlus's website (https://vietnamplus.vn), the website of *Saigon Giai Phong* newspaper (http://sggp.org.vn), website of Thai Nguyen province (http://baothainguyen.org.vn/), the website of *Binh Duong* newspaper (http://baobinhduong.vn/), the website of *Communist* magazine (http://www.tapchicongsan.org.vn/), the website of *Dong Nai* newspaper (http://www.baodongnai.com.vn/), and the website of the Voice of Vietnam (http://vovworld.vn/en-VN.vov). According to the initial survey, the number of Chinese texts on these bilingual websites is limited, and of all eight those of Vietnamplus and the Voice of Vietnam have the most Chinese articles. However, in this paper we do not focus on collecting a bilingual corpus from these websites. We found that there are many Chinese translations of Vietnamese articles on the two websites, but most of them are free translations, and the Vietnamese articles tend to be longer and have more content than the Chinese articles. As such, there will be few truly parallel Chinese-Vietnamese sentence pairs that can be collected if we focus on exploiting these websites.

Therefore, in this paper we propose extracting parallel Chinese-Vietnamese sentence pairs from another source, with the expectation that the extraction will be less time-consuming and the resulting number of parallel sentence pairs will be higher. We choose a multilingual dictionary website (https://glosbe.com/) to collect a Chinese-Vietnamese bilingual corpus. This is a large bilingual resource not only for the Chinese-Vietnamese language pair, but also for many other language pairs. When a Vietnamese word is searched for on the dictionary website, then in addition to returning the Chinese words corresponding to the Vietnamese words, the website also returns many Vietnamese-Chinese example sentence pairs that illustrate the Vietnamese-Chinese or Chinese-Vietnamese word pairs. These bilingual sentence pairs will be a good resource to collect in order to construct a Chinese-Vietnamese bilingual corpus. In short, we chose this dictionary website for this study rather than other bilingual websites for the following reasons:

- In order to extract parallel sentence pairs from bilingual websites, we usually apply the following four steps: (1) downloading the content of Chinese and Vietnamese articles, (2) finding similar Chinese-Vietnamese text pairs (document alignment), (3) finding similar Chinese-Vietnamese sentence pairs from the similar text pairs (sentence alignment), and (4) post-processing. However, these four steps take a lot of time and the quality of the resulting Chinese-Vietnamese bilingual sentence pairs is not very good.
- For the dictionary website, we only need to look up any Vietnamese word and the website will return quite a few bilingual examples to illustrate this. Obviously, the parallelism of these example pairs is better than that of the sentence pairs extracted from other websites.
- The steps to collect bilingual sentence pairs from the dictionary website are much simpler than the four steps given above for extracting from bilingual websites.

The corpus collected from this website will be used for the experiments on the SMT and NMT approaches. We in turn use five bilingual datasets, including 20k, 40k, 60k, 80k, and 100k sentence pairs as training data sets. The initial results indicated that the corpus collected from the Glosbe website showed the following positive results, although there are still some limitations that need to be addressed:

- The data collection method is simple, and can be applied to many other language pairs.
- The quality of the corpus is relatively good with a BLEU score of up to 19.8.
- The non-segmented SMT machine translation gave the best results, and the NMT-segmented translation quality increased monotonously with the amount of training data.

From these results we will propose various ways to improve the quality of the collected data. We will also propose approaches to improve the translation quality of SMT as well as NMT.

The rest of this paper is structured as follows: Section 2 presents some background knowledge of corpus construction, SMT, and NMT. Section 3 presents methods for the construction of a Chinese-Vietnamese bilingual corpus. Section 4 shows and discusses the results of the experiments. Finally, section 5 summarizes our work and gives the main conclusions.

II. BACKGROUND

In this section, we will present some basic knowledge such as the typical corpus collection methods related to Chinese and Vietnamese, along with a brief overview of SMT and NMT.

A. CONSTRUCTION OF BILINGUAL CORPUS

1) BILINGUAL CORPUS CONSTRUCTION FOR RESOURCE-RICH LANGUAGES

Generating a bilingual corpus for resource-rich language pairs such as English, German, Chinese, Japanese, and so on is not too difficult, because the resources for collecting corpora are abundant. Bilingual websites and books are often used to collect these, and with bilingual websites the following four steps are used: (1) downloading content from websites; (2) document alignment; (3) sentence alignment; and (4) post-processing. In addition, some suitable processing steps can be added based on the characteristics of each language pair. The main studies related to such works are briefly reviewed below.

The work in [6] can be seen as the first study on the automatic construction of a Chinese-English bilingual corpus. The author collected more than 10 million Chinese and English texts without specifying the number of collected sentence pairs, and the collection process followed the four-step process outlined above. However, human verification of the alignment results is added in the sentence alignment step, while in the post-processing step the author conducted word segmentation and POS tagging automatically.

In 2009, Liu and Zhou [7] also constructed an English-Chinese bilingual corpus from bilingual websites. In the step of downloading content from the web, the authors analyzed the DOM structure of two websites to improve the efficiency of determining whether the two websites were really parallel or not.

Chu *et al.* [8] used Wikipedia as a resource, and took advantage of the close relationship between Chinese and Japanese to increase the performance of the document alignment and sentence alignment steps. Specifically, they used the shared Chinese characters in both Chinese-Japanese texts as one of the key features to identify similar text/sentence pairs. Like other languages in countries that neighbor China (such as Korea, Vietnam, etc.), Japanese borrows a lot of Chinese characters. In addition, the authors also used a number of other features, such as sentence length, total Chinese characters on each side, the number of words in each other's translations on each side (using a dictionary), and the results of word alignment.

Instead of using a website as a resource to develop a bilingual corpus, Chen and Ge [13] used books to collect an English-Chinese bilingual corpus. The authors took 18 books, including 15 English-Chinese bilingual medical books published by the Shanghai Scientific and Technological Literature Publishing House. The text was then extracted from these books using an OCR scanner to convert to electronic format. The process of building a bilingual corpus was as follows:

- Fix scanner errors: several experienced medical English teachers corrected some mistakes.
- Separate source and target texts.
- Automatic sentence alignment.
- Manual alignment checking.

For the Chinese-English-Chinese machine translation community, the bilingual corpus from Tian *et al.* [1] is extremely helpful. The authors created about 15 million English-Chinese sentence pairs. The corpus has been released to the research community, and is available at the NLP2CT1 website (http://nlp2ct.cis.umac.mo/um-corpus/ index.html). The data sources are online journals, official websites, online language learning resources, the TED website, and microblogs, and so on. The authors also used the four-step process to collect data from bilingual websites. In the post-processing step the authors performed noise filtering to deal with messy codes and mismatched sentence alignments.

The most recent such study is the work of Banon *et al.* [14]. Carried out by a team of 19 members, this study published the ParaCrawl corpus v5.0 consisting of 223 million sentences from around 150k website domains and across 23 EU languages with English. However, the data is highly imbalanced, with 73% of sentences including pairs of just five languages: French, German, Spanish, Italian and Portuguese. However, this is considered the largest bilingual corpus for the research community, and the authors also used the four-step process to construct their individual bilingual corpora.

2) BILINGUAL CORPUS CONSTRUCTION FOR LOW-RESOURCE LANGUAGES

Chinese-Uyghur is a low-resource language pair, and thus it is not easy to create a bilingual corpus because there are not many bilingual electronic documents on the internet. Mamitimin and Dawut [9] initially collected a Chinese-Uyghur bilingual corpus for natural language processing problems. The bilingual text is taken from various sources such as books, newspapers, magazines, and the internet. Non-electronic documents are scanned, OCR applied, and the results reviewed. After this step, all data is saved as text. The collection steps are as follows: Preprocessing → Word/sentence alignment → Manual review and editing → Assign language labels.

Concerning the creation of Vietnamese language corpora, mainly English-Vietnamese-English, we have the work of Ngo *et al.* [15]. The authors created the Vietnamese-English EVB corpora, which consists of both an original English text and its Vietnamese translations, and an original Vietnamese text and its English translations. The original data is from books, including novels and short stories, legal documents, and newspaper articles. The original articles were translated by skilled translators or by contributing authors and were checked again by skilled translators. Each article was translated one-to-one at the whole article level.

In [16], Do *et al.* focused on creating a Vietnamese-French bilingual corpus and using it for SMT. The authors also used some of the common features of the two languages to increase the accuracy of finding similar sentence pairs. Specifically, the authors used features such as: (1) special words: named entities, dates, and numbers; and (2) sentence alignment: sentence length, lexical information. The website http://www.vnagency.com.vn/ was used to extract the bilingual corpus.

In [17], Nguyen *et al.* published a 454K Korean-Vietnamese corpus for machine translation. This corpus was collected from many different data sources: Korean-Vietnamese dictionaries, magazines, books, articles, etc. These data sources were all bilingual websites, well-aligned and well-translated. The construction process also followed the four-step process. In step 4, the authors proposed deleting sentences longer than 80 words, having found that long sentences reduce the quality of machine translation. The experimental results showed that the translation quality is quite good, up to 27.79 BLEU scores, and an automatically extracted bilingual corpus that produces results like this is very impressive.

In 2020, Koehn *et al.* [18] published the bilingual corpora for Pashto-English and Khmer-English language pairs. This was a shared task, and the authors shared a noisy parallel corpus (crawled from the web), and developed methods to align sentences in document pairs and to filter these into a smaller set of high-quality sentence pairs. The authors published many bilingual corpora, of which the biggest are 95,312 sentence pairs for Pashto-English, and 120,156 sentence pairs for Khmer-English. The highest BLEU score is 12.8 for the Pashto-English, and 14.9 for Khmer-English.

Also in 2020, Zhang *et al.* [19] improved machine translation for two low-resource language pairs, Uyghur-Chinese and Mongolian-Chinese. The authors used a Chinese monolingual corpus to learn the vector representations, then applied them to each low-resource language pair. The authors did not mention the process of collecting the Uyghur-Chinese and Mongolian-Chinese bilingual corpora, nor publish them. The authors only indicated that the Uyghur-Chinese corpus has about 170K sentence pairs, and the Mongolian-Chinese corpus contains 260k sentence pairs.

3) CHINESE-VIETNAMESE BILINGUAL CORPUS CONSTRUCTION

With regard to the construction of a Chinese-Vietnamese bilingual corpus, Tran *et al.* [10] is considered the first work to address this issue. In this the authors created about 35,000 Chinese-Vietnamese sentence pairs. However, this corpus was collected manually, which took a lot of time and effort. In 2014, Luo *et al.* [11] also created a Sino-Vietnamese bilingual corpus. However, the authors did not publish the data, so we cannot assess the quality of this corpus.

In [12], Tran *et al.* used movie subtitles as the data source for extraction (http://opus.lingfil.uu.se/OpenSubtitles2016.php). However, this corpus still has many errors such as sentence mismatching, translation errors, free translations, font errors, and so on. The authors thus proposed cleaning up the data as follows: Remove unnecessary symbols → Remove the sentence pairs that contain font errors → Remove the sentences that contain English words → Convert traditional Chinese characters to simplified Chinese characters → Delete sentence pairs with length differences → Delete sentence pairs with very different meanings. However, at present the authors [12] have not yet released their corpus to the research community, so we still do not know its true quality. In addition, we are currently also doing research on a corpus taken from movie subtitles [20], and discovered many misalignment errors.

The most recent work related to the construction of Chinese-Vietnamese bilingual machine translation is [21], which was implemented in 2020 by Li *et al.*. The authors also collected Chinese-Vietnamese bilingual data from news websites, and used the four-step process. In step 3, the authors asked native Vietnamese to edit sentence alignment manually. The final result was 56,610 Chinese-Vietnamese sentence pairs with the highest BLEU score of 16.86. This small corpus has also not been published for the research community.

Unlike the above works, in which the resources are mainly taken from bilingual websites, books, or movie subtitles, our resource for collecting is Glosbe, a multilingual dictionary website. The implementation method used is also simpler and more efficient, as mentioned in Sections 1 and 3.

B. PHRASE-BASED STATISTICAL MACHINE TRANSLATION

Phrase-based Statistical Machine Translation (PSMT) is considered to be the best version of statistical machine translation. This model consists of three components [22] including the phrase translation table $\emptyset\,(\bar{c_i} \mid \bar{v_i})\,\emptyset(\bar{c_i}|\bar{v_i})$, the reordering model R, and the language model P_{LM}. Equation (1) is used to calculate the best translation of a Vietnamese sentence from a Chinese sentence:

$$v_{best} = \emptyset\,(\bar{c_i} \mid \bar{v_i}) * R * P_{LM} \tag{1}$$

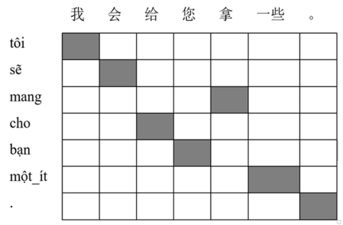
where $(\bar{c_i} \mid \bar{v_i})$ is the pair of Chinese and Vietnamese phrases.

1) THE PHRASE TRANSLATION TABLE

The power of the PSMT model is mainly based on the phrase table (PT). The PT is extracted in two steps, as follows: (1) creation of a word alignment [23] based on a bilingual corpus, and (2) extraction of consistent phrase pairs from this word alignment result. We call a phrase pair $(\bar{c}, \bar{v})$ consistent with an alignment A if all words $c_1, \ldots, c_n$ in $\bar{c}$ that have alignment points in A have these with words $v_1, \ldots, v_n$ in $\bar{v}$ and vice versa:

$$\begin{aligned} &(\bar{c}, \bar{v}) \text{ consistent with } A \\ &\iff \forall v_i \in \bar{v}\colon \left(v_i, c_j\right) \in A \Longrightarrow c_j \in \bar{c} \\ &\quad AND\ \forall c_j \in \bar{c}\colon \left(v_i, c_j\right) \in A \Longrightarrow v_i \in \bar{v} \\ &\quad AND\ \exists v_i \in \bar{v}, c_j \in \bar{c}\colon \left(v_i, c_j\right) \in A \end{aligned} \tag{2}$$

FIGURE 1 illustrates a word alignment result together with the extracted phrase pairs of the sentence pair, “会 给 您 拿 一些 一.” “Tôi sẽ mang cho bạn một_ít.” (I will give you some.)



tôi – 我
tôi sẽ – 我会
tôi sẽ mang cho bạn – 我 会 给 您 拿
tôi sẽ mang cho bạn một_ít – 我 会 给 您 拿一些
tôi sẽ mang cho bạn một_ít . – 我 会 给 您 拿一些 。
sẽ –会
sẽ mang cho bạn – 会 给 您 拿
sẽ mang cho bạn một_ít – 会 给 您 拿一些
sẽ mang cho bạn một_ít . – 会 给 您 拿一些 。

mang –拿
mang cho bạn – 给 您 拿
mang cho bạn một_ít – 会 给 您 拿一些
mang cho bạn một_ít . – 会 给 您 拿一些 。
cho – 给
cho bạn – 给 您
bạn – 您
một_ít –一些
một_ít . –一些 。

FIGURE 1. An example of extracted phrase pairs from the word alignment process.

2) REORDERING MODEL

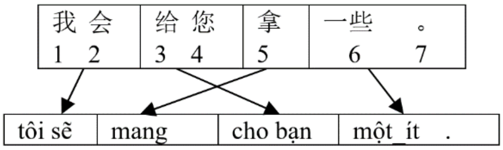
The distance-based reordering model (DRM) and the lexicalized reordering model (LRM) are two modes that are commonly used in PSMT. The DRM is defined as follows:

$$d\,(x) = \alpha^{|x|}, \quad \alpha \in [0, 1] \tag{3}$$

where $\mathrm{x} = \mathrm{start_i} - \mathrm{end_{i-1}} - 1$. The term $\mathrm{start_i}$ is the position of the first word in the input Chinese phrase that is translated into the i-th word of the Vietnamese phrase. The term end$_i$ is the position of the last word of that Chinese phrase.

FIGURE 2 illustrates the distance of the word order movement of the sentence pair "我 会 给 您 拿 一些" → "Tôi sẽ mang cho bạn một_ít." (I will bring you some).

The probability of the DRM is inversely proportional to the distance of the two phrases. This model is effective for language pairs that are less reordered or have a short distance (such as Arabic-English or French-English) due to the probability of the P_{LM} language model compensating for the cost of reordering. However, this model is not good for language pairs that have a long distance. Some long-distance language pairs are Japanese-English and Chinese-Vietnamese, among other examples.



Phrase	Translates	Movement	Distance
1	1-2	Start at begining	0
2	5	skip over 3-4	+2
3	3-4	move back over 3-5	-3
4	6-7	skip over 5	+1

FIGURE 2. An example of distance-based reordering.

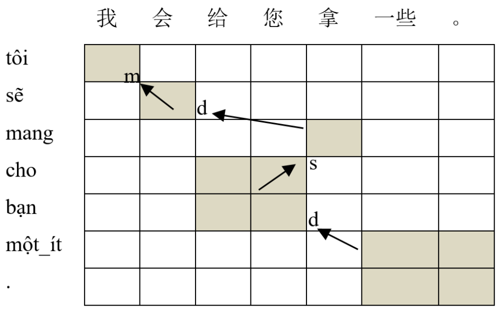


FIGURE 3. Three different orientations of a phrase in the lexicalized reordering model.

The LRM is concerned with three types of word order changes, including monotone (m), swap (s), and discontinuous (d). The three types are learned from the word alignment results. Equation (4) illustrates the probabilities of these three types and FIGURE 3 shows an example of three different orientations.

$$orientation \in \{m, s, d\};$$
$$p_o\left(orientation\,|\bar{f},\bar{e}\right) = \frac{count(orientation,\bar{e},\bar{f})}{\sum_o count(o,\bar{e},\bar{f})} \qquad (4)$$

3) LANGUAGE MODEL

The language model helps a translation system determine the accuracy of word order in the generated sentence. Using the generated word sequence, the translation system computes the frequency of these words in the target language. This information is used in the decoding process of PSMT to find the best translation of a sentence or phrase. The N-gram language model is often used in state-of-the-art PSMT.

Equation 5 presents the bi-gram language model. In which w1, w2 are two words in a sentence $s = w_1, w_2, \ldots, w_n$:

Bi-gram probability:

$$P\left(w_2|w_1\right) = \frac{count(w_1w_2)}{count(w_1)} \qquad (5)$$

The occurrence probability of the sentence s is calculated by the product of all probabilities of each word contained in the sentence s. Here is an example of how the occurrence probability of the sentence "I go to school" is determined in the bi-gram language model. The symbols <s> and </s> indicate the beginning and ending of the sentence, respectively. For instance, $P(I\ go\ to\ school) = P(I|\ \ <s>) \times P(go|I) \times P(to|go) \times P(school|to) \times P(</s>\ |school)$.

C. NMT

Currently, the state of the art in NMT is the sequence-to-sequence encoder-decoder model. This model uses an end-to-end mechanism including two parts: the encoder and decoder. The encoder and decoder contain one or more neural networks. Recurrent neural networks (RNNs), such as long-short-term memory (LSTM) or gated recurrent units (GRUs), are most suitable for NMT to deal with input sentences of various lengths. Moreover, NMT also takes advantage of attention mechanisms in guiding the decoder to determine which part of the encoding is relevant at each step of the generation. In this subsection, we introduce the fundamental architecture of an NMT, including an encoder, decoder, and attention mechanism.

1) ENCODER

The encoder module contains basically one or more RNNs, which are responsible for encoding a representation of an input sentence. The input sentence is a sequence of words, which is encoded into word representation vectors or embedding vectors. These vectors represent words in a continuous space, and subsequently the vectors are pushed into RNNs, resulting in hidden states. An encoder usually has one RNN encoding words from left to right and one RNN encoding words from right to left. This combination is called a bidirectional RNN.

FIGURE 4 illustrates the encoder module. This encoder contains the embedding representation for each input word x_j, as well as the combination of forward RNN and backward RNN ($[\overrightarrow{h_j}, \overleftarrow{h_j}]$).

$$\overrightarrow{h_j} = RNN\left(\overrightarrow{h_{j+1}}, \bar{E}x_j\right); \quad \overleftarrow{h_j} = RNN(\overleftarrow{h_{j+1}}, \bar{E}x_j) \qquad (6)$$

Any RNN in an encoder can be either LSTM or GRU. Approaches with GRU are more current, but LSTM units are widely used.

2) DECODER

Because the output of a decoder is also a sequence of words, RNNs are also used in this. These RNNs take input context

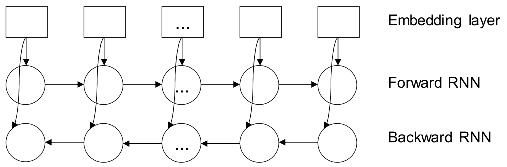


FIGURE 4. An illustration of the NMT encoder.

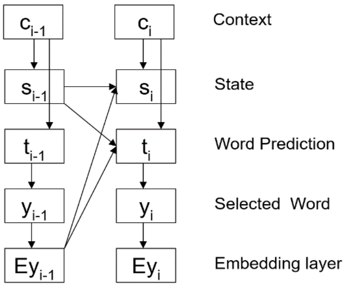


FIGURE 5. Training steps in the NMT decoder.

from the encoder, from the preceding hidden state, and from the previous output word prediction, to generate a new hidden decoder state and a new output word prediction.

FIGURE 5 shows the components of a decoder. If the decoder uses an attention mechanism, the context vector is the output of the attention layer. Otherwise, the context vector is simply the combination of $[\overrightarrow{h_j}, \overleftarrow{h_j}]$.

$$s_i = RNN\left(s_{i-1}, Ey_{i-1}, c_i\right) \tag{7}$$

In the hidden states s_i, the decoder calculates the probability distribution for all known words by using a softmax function. For example, if the number of all known words is 10,000, then the output of this softmax layer is a 10,000-dimensional vector corresponding to each word's probability.

The prediction vector t_i is calculated by the decoder hidden state s_{i-1} and the embedding of the previous output word Ey_{i-1} and the input context c_i.

$$t_i = softmax(W\left(Us_{i-1} + VEy_{i-1} + Cc_i\right)) \tag{8}$$

In the training phase, the correct output y_i is known. The cost function is the negative *log*, which is defined as:

$$C = -log t_i[y_i] \tag{9}$$

In the inference phase, it is simple to choose the word with the highest probability at each generation step. However, the combination of all the words with the highest probabilities is not always the best-translated sentence. As a solution, the beam search, or greedy search, is used to select the best output sentence.

3) ATTENTION MECHANISM

The attention mechanism comes between the encoder and decoder to help the latter determine which encoder inputs are more important at each step of the decoding process. FIGURE 6 shows an example of the attention layer.

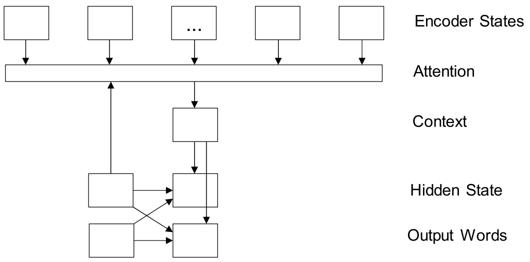


FIGURE 6. An examination of the attention layer.

There are currently two popular publications for the attention mechanism, which are Luong *et al.* [24] and Bahdanau *et al.* [25].

The approach proposed by Luong *et al.* [24] consists of the following stages:

- The current target hidden state is calculated with all source states to get attention weights.

$$a_{ts} = softmax(score(h_t, \overline{h}_s)) \tag{10}$$

- The context vector is computed as the weighted average of the source states.

$$c_t = \sum\nolimits_s a_{ts}\overline{h_s} \tag{11}$$

- The context vector is combined with the current target hidden state to have the attention vector.

$$a_t = tanh(W_c[c_t; h_t]) \tag{12}$$

Currently, the score in Equation 12 is calculated by a method proposed by Luong *et al.* [24] and Bahdanau *et al.* [25].

Luong's multiplicative style

$$score\left(h_t, \overline{h}_s\right) = h_t^T W\overline{h}_s \tag{13}$$

Bahdanau's additive style

$$score\left(h_t, \overline{h}_s\right) = v_a^T tanh(W_1 h_t + W_2\overline{h}_s) \tag{14}$$

III. CONSTRUCTION OF A BILINGUAL CORPUS FROM THE GLOSBE WEBSITE

A. CONSTRUCTION METHOD

Our construction method also follows the four-step process of corpus construction from a bilingual website. However, we skip Step 2 (Document alignment) and Step 3 (Sentence alignment), focusing only on Step 1 (Downloading content) and Step 4 (Post-processing). Steps 2 and 3 are not necessary because the results returned when looking up a word on the Glosbe website already include pairs of example sentences that can be considered translations of each other. FIGURE 7 presents a method to construct a Chinese-Vietnamese bilingual corpus. This method is applicable to any other language pair supported by Glosbe, especially language pairs where one of the two languages has a close relationship with Chinese such as Korean, Japanese, etc.

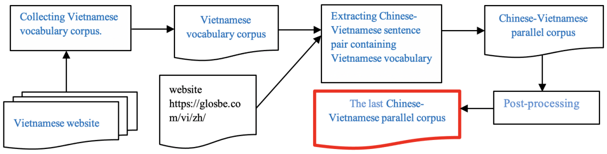
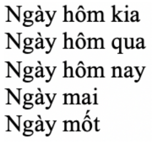

Collecting Vietnamese vocabulary corpus.
Vietnamese vocabulary corpus
Extracting Chinese-Vietnamese sentence pair containing Vietnamese vocabulary
Chinese-Vietnamese parallel corpus
Vietnamese website
website https://glosbe.com/vi/zh/
The last Chinese-Vietnamese parallel corpus
Post-processing

FIGURE 7. Steps to extract the Chinese-Vietnamese bilingual corpus.

B. COLLECTING VIETNAMESE VOCABULARY

A Vietnamese vocabulary corpus is used to look up the Glosbe dictionary. We collected this vocabulary corpus from two different sources: (1) from Wikipedia's Vietnamese language category,[1] and (2) from the *Saigon Giai Phong* newspaper website.[2] The vocabulary on Wikipedia is similar to that in a Vietnamese dictionary. We do not use an existing Vietnamese dictionary because of copyright issues. From Wikipedia, the system has collected almost all of the existing Vietnamese vocabulary. Next, the system adds words from the website (2) to the Vietnamese vocabulary. We use this because it is a news website so the content is updated regularly, and there are some new words that may not be in Wikipedia. An example from the collected Vietnamese vocabulary corpus is shown in FIGURE 8.



Ngày hôm kia
Ngày hôm qua
Ngày hôm nay
Ngày mai
Ngày mốt

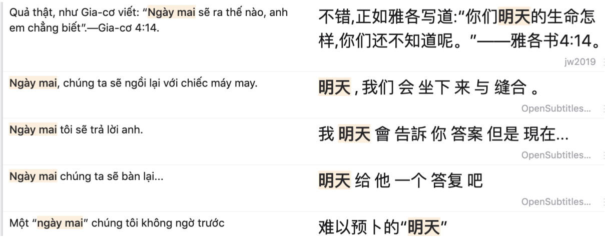
FIGURE 8. An example from the Vietnamese vocabulary corpus.

C. EXTRACTING CHINESE-VIETNAMESE SENTENCE PAIRS CONTAINING VIETNAMESE VOCABULARY

The steps to extract the Chinese-Vietnamese sentence pairs are as follows:

- For each Vietnamese word v in the "Vietnamese vocabulary corpus", the system will check the Chinese meaning of the word v at the address: "https://glosbe.com/vi/zh/" $+v$.
- Glosbe will return the Chinese meaning of the Vietnamese vocabulary v. At the same time, many Chinese-Vietnamese sentence pairs illustrating the Vietnamese word v *will* be returned. FIGURE 9 illustrates the results of Glosbe when looking up the Vietnamese $v =$ "ngày mai" (tomorrow).
- The extraction system only extracts Chinese-Vietnamese examples sentence pairs and omits other contents.

Glosbe returns bilingual sentence pairs at the sentence level, not the document level like other web resources. Therefore, there is no document or paragraph alignment step in our method. As such, we do not use common sentence alignment methods like yalign[3] or hunalign[4] to extract parallel sentence pairs from parallel documents or paragraphs.



Quả thật, như Gia-cơ viết: "Ngày mai sẽ ra thế nào, anh em chẳng biết".—Gia-cơ 4:14.	不错,正如雅各写道:"你们明天的生命怎样,你们还不知道呢。"——雅各书4:14。 jw2019
Ngày mai, chúng ta sẽ ngồi lại với chiếc máy may.	明天 , 我们 会 坐下 来 与 缝合 。 OpenSubtitles...
Ngày mai tôi sẽ trả lời anh.	我 明天 會 告訴 你 答案 但是 現在... OpenSubtitles...
Ngày mai chúng ta sẽ bàn lại...	明天 给 他 一个 答复 吧 OpenSubtitles...
Một "ngày mai" chúng tôi không ngờ trước	难以预卜的"明天"

FIGURE 9. The results returned from Glosbe when looking up the Vietnamese vocabulary item v = "ngày mai" (accessed June 7, 2021).

D. THE ORIGINAL CORPUS

The original corpus consists of 740,691 sentence pairs. However, this corpus still has many errors that need to be corrected. Some common errors are as follows:

- There are quite a lot of garbage words in both Chinese and Vietnamese sentences.
- Sentence pairs are not really parallel: Chinese sentences contain English translations, Chinese and Vietnamese sentence differ greatly in length.
- There are quite a few duplicate sentences.

We do post-processing to increase the quality of the corpus.

E. POST-PROCESSING

In this step we will conduct post-processing of the corpus, mainly carrying out three tasks: deleting garbage words, deleting sentence pairs that are not really parallel, and deleting duplicate sentences.

1) DELETING GARBAGE WORDS

The sentence pairs from the Glosbe website are taken from a variety of sources, and the sentence pairs in each source often have garbage words attached to them. For example, the sentence pair taken from the TED website will have the garbage word "TED" at the end of the sentence, or the sentence pair taken from the movie subtitles will have the garbage word "OpenSubtitles..." at the end of the sentence. These garbage words have no meaning for a bilingual sentence pair, so in this step the system will remove these from the bilingual corpus. From our experiments we found some common garbage words in Chinese-Vietnamese sentence pairs, such as "JW_2017_12", "TED", "LDS", "OpenSubtitles2018".

2) DELETING SENTENCE PAIRS THAT ARE NOT REALLY PARALLEL

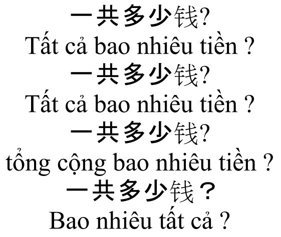
There are quite a few sentence pairs that are not really parallel in the original corpus, specifically: Chinese sentences have English translations, Vietnamese translations only partially match Chinese sentences, etc. We use the length factor and Sino-Vietnamese words of Chinese-Vietnamese sentence pairs to filter out the Chinese-Vietnamese sentence pairs that are not really parallel. According to [10], the length difference

[1] Download at: https://vi.wiktionary.org/w/index.php?title=Th%E1%BB%83_lo%E1%BA%A1i:Danh_t%E1%BB%AB_ti%E1% BA%BFng_Vi%E1%BB%87t

[2] Download at: http://www.sggp.org.vn/

[3] https://pythonrepo.com/repo/machinalis-yalign-python-natural-language-processing

[4] http://mokk.bme.hu/en/resources/hunalign/



一共多少钱?
Tất cả bao nhiêu tiền ?
一共多少钱?
Tất cả bao nhiêu tiền ?
一共多少钱?
tổng cộng bao nhiêu tiền ?
一共多少钱？
Bao nhiêu tất cả ?

FIGURE 10. Illustration of the duplicate Chinese sentence "一共多少钱" in the collected bilingual corpus.

between Chinese and Vietnamese sentences is about 10%. In this work, we only keep the Chinese-Vietnamese sentence pairs whose length ratio differs by not less than 30% or there are Sino-Vietnamese translations in both Chinese and Vietnamese sentences. According to [10], Vietnamese borrows about 65% of Chinese vocabulary and is called Sino-Vietnamese. We use Sino-Vietnamese dictionary to identify Sino-Vietnamese words in Chinese and Vietnamese sentences. Sentence pairs that have a length ratio greater than 30% but have Sino-Vietnamese translations on both sides will be kept.

3) DELETING DUPLICATE SENTENCES

Empirically, we found that Glosbe uses the same sentence pair to illustrate many different Vietnamese vocabulary items. For example, "一共多少钱 ?" → "Tất cả bao nhiêu tiền ?" (how much is the total) is used for many illustrative examples, including "tất cả" (the total), "bao nhiêu" (how much), "tiền" (money). In the process of browsing all the vocabulary in the "Vietnamese vocabulary corpus" to search on the Glosbe page, the system will get many duplicate sentence pairs. In this step, the system will thus filter the duplicate (or close to duplicate) sentences and keep only one sentence. We considered the sentences duplicate when their cosine measurement was 95% or more. FIGURE 10 shows some duplicate sentences in the corpus.

F. THE FINAL CHINESE-VIETNAMESE BILINGUAL CORPUS

We collected the Chinese-Vietnamese corpus from Glosbe with a total of 740,691 sentence pairs. After removing the duplicates, the remaining corpus was 428,639 sentence pairs. We extracted 100,000 sentence pairs from 428,639 pairs of these sentences to conduct experiments for PSMT and NMT. We continued to subdivide these 100k sentence pairs into five corpora, including 20k (GL20), 40k (GL40), 60k (GL60), 80k (GL80) and 100k (GL100) sentence pairs. Corpora are extracted according to the rule that a larger corpus will consist of a smaller corpus. For example, the GL40 corpus already includes the sentence pairs of GL20.

We found that the Glosbe corpus was taken from a variety of resources, such as: OpenSubtitle, Tatoeba, Bible, Wikipedia, TED, etc. For each resource, the style or length of the example sentence pairs will be different. For example, sentences from OpenSubtitle or Tatoeba tend to be conversational style in everyday life, in movies and their length is usually short; while sentence pairs taken from the Bible will be longer and have a religious style. Therefore, the Glosbe corpus includes both simple sentences of short length and complex sentences of long length.

TABLE 1. Statistics about the GL100 Corpus.

Corpora	BS-Chinese	BS-Vietnamese	WS-Chinese	WS-Vietnamese
NS	90,000	90,000	90,000	90,000
NW	1,157,839	1,197,833	779,197	958,954
NW/NS	12.86	13.30	8.66	10.66
MIN LEN	3	4	2	4
MAX LEN	123	129	76	102
Vocabulary size	5,148	13,142	39,672	24,196

Table 1 illustrates some statistics about training corpus of the GL100. In which,

- NS is the number of sentences.
- NW is the number of words.
- MIN LEN is the shortest length of a sentence.
- MAX LEN is the longest length of a sentence.
- BS-Chinese: The Chinese corpus is segmented at the character level.
- BS-Vietnamese: The Vietnamese corpus is segmented at the syllable level.
- WS-Chinese: The Chinese corpus is segmented at the word level.
- WS-Vietnamese: The Chinese corpus is segmented at the word level.

IV. COMPARISION OF SMT AND NMT IN CHINESE-VIETNAMESE MACHINE TRANSLATION

A. TOOLKITS

In the Chinese and Vietnamese languages, the words written in a sentence are not broken up by a space. Thus in Chinese or Vietnamese machine translation, the word segmentation problem is often resolved first, before further rendering into another language. We used the Stanford Segmenter [26] to spatially distinguish words in the Chinese corpus.[5] As for Vietnamese, we used the CLC_VN_WS toolkit [27] to segment words.[6] In addition, we used the TensorFlow-based attention NMT model[7] for training and testing our Chinese-Vietnamese NMT systems. For PSMT training and testing, the state-of-the-art Moses toolkit[8] was used. We used BLEU scores [28] to evaluate the quality of translation. Table 2 presents the parameters in the training model of SMT and NMT.

[5] https://nlp.stanford.edu/software/segmenter.html
[6] http://www.clc.hcmus.edu.vn/?page_id=36
[7] https://github.com/tensorflow/nmt/
[8] http://www2.statmt.org/moses/?n=Moses.Baseline

TABLE 2. The parameters in SMT and NMT training models.

MT Models	Parameters	Values
SMT	-alignment -reordering -lm	grow-diag-final msd-bidirectinal-fe 0:3
NMT	Optimizer lr clip-norm dropout max-tokens label-smoothing lr-scheduler arch	Nag 0.5 0.1 0.2 4,000 0.1 Fixed Fconv

B. EXPERIMENTAL CORPORA

We use five bilingual corpora including GL20, GL40, GL60, GL80 and GL100 to conduct machine translation experiments.[9] We used 90% of the sentences for training, 5% for testing, and the remaining 5% for developing. We divided each corpus into three components, as follows: in every 20 sentences the first to eighteenth sentences were used for training, the nineteenth sentence for developing, and the twentieth sentence for testing.

For each corpus we performed four experiments, including Baseline NMT translation (BS-NMT), Word segmentation NMT translation (WS-NMT), Baseline PSMT translation (BS-PSMT), and Word segmentation PSMT translation (WS-PSMT).

- BS-NMT and BS-PSMT: Chinese characters and Vietnamese words are seen as independent units. One space is inserted between Chinese characters. For Vietnamese, one space is inserted between words and punctuation.
- WS-NMT and WS-PSMT: Words in the Chinese corpora are segmented by the Chinese Segmenter. For Vietnamese, words are segmented using the CLC_VN_WS toolkit.

In addition, we also used the OpenSubtitle2016 corpus to conduct experiments. The purpose of comparing Glosbe and OpenSubtile corpora is to highlight the post-processing of our proposed method. This comparison is not intended to show the quality of the two corpora because inherently the number of sentence pairs in the two corpora is different, the domain of the two corpora is also different. This dataset includes 1,076,805 sentence pairs, of which we use 985,288 sentence pairs for training, 46,732 sentence pairs for developing, and 44,785 sentence pairs for testing. Table 3 displays statistics about the OpenSubtitle2016 corpus. Due to the nature of the conversational text, the word-to-sentence ratio (NW/NS) of OpenSubtitle2016 corpus is lower than that of Glosbe corpus.

TABLE 3. Statistics about the OpenSubtitle2016 corpus.

Corpora	BS-Chinese	BS-Vietnamese	WS-Chinese	WS-Vietnamese
NS	956,171	956,171	956,171	956,171
NW	10,053,462	9,526,722	7,940,196	8,643,163
NW/NS	10.51	9.96	8.30	9.03
MIN LEN	1	1	1	1
MAX LEN	118	185	112	100
Vocabulary size	13,299	8,983	15,235	39,210

TABLE 4. Experimental results.

Corpora	BS-NMT	BS-SMT	WS-NMT	WS-SMT
GL20	15.34	**15.77**	14.62	12.77
GL40	16.08	**17.5**	15.9	14.28
GL60	18.01	**18.9**	18.11	15.5
GL80	18.51	**19.18**	18.12	15.99
GL100	18.89	**19.8**	19.38	16.93
OpenSubtitle 2016	11.76	**12.2**	12.14	11.21

C. EXPERIMENTAL RESULTS

Table 4 shows the BLEU scores of the system's bilingual corpora.

D. DISCUSSION

From the experiments that are reported in the Table 4, we found that: (1) the quality of the BS-SMT translation systems is the best of all cases, (2) WS-NMT gives better results than BS-NMT when the training data is gradually increased. (3) the larger the training data, the higher the translation quality, (4) the quality of the corpus needs to be improved, and (5) translation quality on the OpenSubtitle2016 corpus is not as good as our corpus.

1) THE QUALITY OF BS-SMT TRANSLATION SYSTEMS IS THE BEST OF ALL CASES

This is an interesting result, because we know that in recent studies the quality of NMT translation has been better than that of SMT translation, especially in resource-rich

[9] Readers who wish to use these corpora for research are asked to please contact us through following the email address: tranthanhphuoc@tdtu.edu.vn or gocong06@gmail.com

TABLE 5. **A sentence pair of the GL100 corpus.**

Chinese sentence C	为什么 研读 圣经 应该 留意 基督徒 的 道德 标准 ?
Vietnamese sentence V	Tại sao khi học Lời Đức Chúa Trời , bạn nên lưu ý đến các nguyên tắc đạo đức ?
English meaning of V	Why should you pay attention to moral principles when studying God's Word?
The best Vietnamese meaning V'	Tại sao việc học Kinh Thánh phải chú ý đến các tiêu chuẩn đạo đức của Cơ đốc giáo?
English meaning of V'	Why should studying the Bible pay attention to Christian moral standards ?

languages [29]–[32]. With very few language pairs does SMT give better results than NMT, with only one work to date showing this [33] (for Arabic-English language pairs). We think that it is necessary to do more experiments on NMT and SMT to find out the pros and cons of the two translation systems, so that we can integrate the advantages of both as one of the possible research directions in the near future.

2) WS-NMT GIVES BETTER RESULTS THAN BS-NMT WHEN THE TRAINING DATA IS GRADUALLY INCREASED

Chinese and Vietnamese are the same language type (isolated language) [10], where the words are not distinguished by a space. Therefore, the word segmentation problem is often resolved first in Chinese or Vietnamese machine translation before further rendering into another language.

In [10], Tran *et al.* tested the Chinese-Vietnamese SMT for the cases of word segmentation and non-segmentation on a good quality corpus (due to manual collection). The results showed that the translation quality of these two cases was almost the same, with WS giving higher BLEU scores three times and BS two times. The difference in the BLEU scores of the five experiments was also not large, the highest was 0.35 (WS was 34.87 and BS was 34.52) and the lowest was 0.02 (BS was 34.81 and WS was 34.79). However, from the experiments carried out in the current work we found that for NMT word segmentation translation will give better results than non-segmentation translation.

3) THE BIGGER THE TRAINING DATA, THE HIGHER THE QUALITY OF THE TRANSLATION

This issue is obvious for machine translation, whether SMT or NMT. The larger the corpus is, the more cases the system will learn, and the better the translation results will be.

4) THE QUALITY OF THE COLLECTED CORPUS NEEDS TO BE IMPROVED

Indeed, we found that with a corpus of 100k sentence pairs, the highest BLEU score of 19.8 is not high, and obviously this corpus needs to be improved. From the experiments we found that the corpus has some limitations, as follows:

- Because sentence pairs are examples for a certain word, the content of the corpus is often discrete, and the sentences in the corpus are often unrelated to each other.
- At the time of collecting the corpus there were quite a few example sentence pairs on the website related to religion, specifically Christianity. Therefore, the style in the corpus has a lot of religious features.
- The corpus is automatically collected, so there are still some errors, such as free translation, multi-meaning translation (one Chinese sentence translates into many Vietnamese sentences, usually separated by; or ()). In the near future, we will study and apply some automatic/semi-automatic data cleaning methods to increase the quality of this corpus.

Table 5 presents an example of a sentence pair in the GL100 corpus. The Vietnamese sentence V is a translation of the Chinese sentence C in the corpus. The Vietnamese sentence V' is the best translation of the Chinese sentence C. The word "圣经" translates to "the Bible" (translation V') better than "God's Word" (translation V). The word "标准" is a Sino-Vietnamese that means "standard" (translation V'), not "rule" (translation V). "基督徒" means Christianity, the V translation omits this word.

5) THE QUALITY OF OpenSubtitle2016'S CORPUS IS NOT AS GOOD AS OURS

The OpenSubtitle2016 corpus has more than 1 million sentence pairs (10 times more than our corpus), but the BLEU score is much lower. We surveyed this corpus and found that its quality is not really good, and there are quite a few mis-alignment errors. Here is a case of mis-alignment of the OpenSubtitle2016 corpus: Chinese sentence: 你可以不相信我 You don't have to believe me.". Its Vietnamese translation is "Cô nói úng, không cần phả i tin tôi" (English meanning: "You're right, you don't have to believe in me"). In this case, the Chinese side of the sentence contains its English translation. Cases like this are automatically removed from our corpus.

V. CONCLUSION

In this paper, we have presented a method to collect a Chinese-Vietnamese bilingual corpus from the Glosbe multilingual dictionary website, and have collected more than 400k Chinese-Vietnamese bilingual sentence pairs from this website. We will publish 100k sentence pairs from the obtained corpus for the natural language processing community, and especially the machine translation community, for research purposes. The collection of data from this dictionary website is rapid, and the number of sentence pairs collected in a short period of time is considerable. Moreover, this collection method can be applied to many other language pairs, not just Chinese-Vietnamese. In the near future, we plan to apply this method to create a multilingual corpus, focusing on languages of countries near Vietnam, such as ASEAN countries.

However, besides these advantages the collected corpus also has some limitations that we will work to overcome in the near future.

In addition to automatically collecting a Chinese-Vietnamese bilingual corpus from the Glosbe dictionary website, it is necessary to study methods of collection from other resources, especially Chinese-Vietnamese bilingual websites. Compared to dictionary websites, bilingual ones have more updated and richer information. Therefore, one of the next tasks that we carry out is to continue researching and building a Chinese-Vietnamese bilingual corpus from bilingual websites.

Through the experiments examining two translation systems, NMT and SMT, we found that for a low-resource language pair, and when spaces do not define word boundaries, like in Chinese or Vietnamese, the SMT system with word non-segmentation still gives the best results. However, the NMT system with word segmentation gives better results when the corpus is larger. The integration of segmentation and non-segmentation factors as well as the advantages of SMT and NMT are essential to improve the quality of machine translation in the future.

REFERENCES

[1] L. Tian, D. F. Wong, L. S. Chao, P. Quaresma, F. Oliveira, Y. Lu, S. Li, Y. Wang, and L. Wang, "UM-corpus: A large English–Chinese parallel corpus for statistical machine translation," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 1837–1842.

[2] *German-English Parallel Corpus 'de-news', Daily News 1996–2000*. Accessed: Oct. 4, 2021. [Online]. Available: https://homepages.inf.ed.ac.uk/pkoehn/publications/de-news/

[3] *German-English Website Parallel Corpus*. Accessed: Oct. 4, 2021. [Online]. Available: https://data.europa.eu/data/datasets/elrc_41?locale=en

[4] Q. H. Ngo, D. Dinh, and W. Winiwarter, "Automatic searching for English–Vietnamese documents on the internet," in *Proc. 3rd Workshop South Southeast Asian Natural Lang. Process. (SANLP)*, 2012, pp. 211–220.

[5] Q. H. Ngo and W. Winiwarter, "Building an English–Vietnamese bilingual corpus for machine translation," in *Proc. Int. Conf. Asian Lang. Process.*, Nov. 2012, pp. 157–160, doi: 10.1109/IALP.2012.30.

[6] B. B. Chang, "Chinese–English parallel corpus construction and its application," in *Proc. 18th Pacific Asia Conf. Lang., Inf. Comput.*, 2004, pp. 283–290.

[7] D.-F. Liu and X. Zhou, "A method of construction of the Chinese and English bilingual translation corpus based on web data mining," in *Proc. Asia–Pacific Conf. Inf. Process.*, Jul. 2009, pp. 317–319, doi: 10.1109/APCIP.2009.87.

[8] C. H. Chu, T. Nakazawa, and S. Kurohashi, "Constructing a Chinese–Japanese parallel corpus from Wikipedia," in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014, pp. 642–647.

[9] S. Mamitimin and U. Dawut, "Chinese–Uyghur parallel corpus construction and its application," in *Proceedings of Using Corpora in Contrastive and Translation Studies*. Cambridge Scholars, 2008, pp. 281–295.

[10] P. Tran, D. Dinh, and L. H. B. Nguyen, "Word re-segmentation in Chinese–Vietnamese machine translation," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 16, no. 2, pp. 1–22, Dec. 2016, doi: 10.1145/2988237.

[11] L. Luo, J.-y. Guo, Zheng- tao Yu, Y.-y. Mo, and L.-J. Zhou, "Construction of a large-scale Sino–Vietnamese bilingual parallel corpus," in *Proc. IEEE Int. Conf. Syst. Sci. Eng. (ICSSE)*, Jul. 2014, pp. 154–157, doi: 10.1109/ICSSE.2014.6887924.

[12] H. A. Tran, Y. H. Guo, P. Jian, S. M. Shi, and H. Y. Huang, "Improving parallel corpus quality for Chinese–Vietnamese statistical machine translation," *J. Beijing Inst. Technol.*, vol. 27, no. 1, pp. 127–136, 2018.

[13] X. X. Chen and S. L. Ge, "The construction of English–Chinese parallel corpus of medical works based on self-coded Python programs," in *Proc. Int. Conf. Adv. Eng.*, 2011, pp. 598–603.

[14] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz-Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza, "ParaCrawl: Web-scale acquisition of parallel corpora," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4555–4567.

[15] Q. H. Ngo, W. Winiwarter, and B. Wloka, "EVBCorpus—A multi-layer English–Vietnamese Bilingual corpus for studying tasks in comparative linguistics," in *Proc. Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 14–18.

[16] T. N. D. Do, V. B. Le, B. Bigi, L. Besacier, and E. Castelli, "Mining a comparable text corpus for a Vietnamese—French statistical machine translation system," in *Proc. 4th EACL Workshop Stat. Mach. Transl.*, 2009, pp. 165–172.

[17] Q.-P. Nguyen, A.-D. Vo, J.-C. Shin, P. Tran, and C.-Y. Ock, "Korean–Vietnamese neural machine translation system with Korean morphological analysis and word sense disambiguation," *IEEE Access*, vol. 7, pp. 32602–32616, 2019, doi: 10.1109/ACCESS.2019.2902270.

[18] P. Koehn, V. Chaudhary, A. El-Kishky, N. Goyal, P. J. Chen, and F. Guzman, "Findings of the WMT 2020 shared task on parallel corpus filtering and alignment," in *Proc. 5th Conf. Mach. Transl. (WMT)*, 2020, pp. 726–742.

[19] X. Zhang, X. Li, Y. Yang, and R. Dong, "Improving low-resource neural machine translation with teacher-free knowledge distillation," *IEEE Access*, vol. 8, pp. 206638–206645, 2020, doi: 10.1109/ACCESS.2020.3037821.

[20] *OpenSubtitles 2016*. Accessed: Oct. 5, 2021. [Online]. Available: https://opus.nlpl.eu/download.php?f=OpenSubtitles/v2016/tmx/vi-zh.tmx.gz

[21] H. Li, J. Sha, and C. Shi, "Revisiting back-translation for low-resource machine translation between Chinese and Vietnamese," *IEEE Access*, vol. 8, pp. 119931–119939, 2020, doi: 10.1109/ACCESS.2020.3006129.

[22] P. Koehn, A. Axelrod, A. B. Mayne, C. C. Burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proc. 2nd Int. Workshop Spoken Lang. Transl. (IWSLT)*, Pittsburgh, PA, USA, 2005, pp. 68–75.

[23] A. Fraser and D. Marcu, "Measuring word alignment quality for statistical machine translation," *Comput. Linguistics*, vol. 22, no. 3, pp. 293–303, Sep. 2007, doi: 10.1162/coli.2007.33.3.293.

[24] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 1412–1421.

[25] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[26] H. Tseng, P. C. Chang, G. Andrew, D. Jurafsky, and C. Manning, "A conditional random field word segmenter for Sighan Bakeoff 2005," in *Proc. 4th SIGHAN Workshop Chin. Lang. Process.*, Jeju, South Korea, 2005, pp. 168–171.

[27] D. Dinh and T. Vu, "A maximum entropy approach for Vietnamese word segmentation," in *Proc. Int. Conf. Res., Innov. Vis. Future (RIVF)*. Hanoi, Vietna: HCMC, 2006, pp. 247–252.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.

[29] S. Kinoshita, T. Oshio, and T. Mitsuhashi, "Comparison of SMT and NMT trained with large patent corpora: Japio at WAT2017," in *Proc. 4th Workshop Asian Transl. (WAT)*, Taipei, Taiwan, 2017, pp. 140–145.

[30] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, "Is neural machine translation ready for deployment? A case study on 30 translation directions," 2016, *arXiv:1610.01108*.

[31] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: A case study," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 257–267.

[32] M. Popović, "Comparing language related issues for NMT and PBMT between German and English," *Prague Bull. Math. Linguistics*, vol. 108, pp. 209–220, Jun. 2017.

[33] M. A. Menacer, D. Langlois, O. Mella, D. Fohr, D. Jouvet, and K. Smaïli, "Is statistical machine translation approach dead?" in *Proc. Int. Conf. Natural Lang., Signal Speech Process. (ICNLSSP)*, Casablanca, Morocco, 2017, pp. 1–5.

• • •