ICIC Express Letters
Part B: Applications
Volume **16**, Number **7**, July **2025**

 ISSN 2185-2766
pp. **727–734**

BUILDING CHINESE-VIETNAMESE BILINGUAL CORPUS FROM TED-TALKS

Phuoc Tran[1], Phuc-Nghi Nguyen[1] and Dinh-Tu Truong[2,*]

[1]Natural Language Processing and Knowledge Discovery Laboratory
Faculty of Information Technology
Ton Duc Thang University
19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City 700000, Vietnam
{ tranthanhphuoc; tg_nguyenphucnghi }@tdtu.edu.vn

[2]Faculty of Information Technology
Ton Duc Thang University
19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City 700000, Vietnam
*Corresponding author: truongdinhtu@tdtu.edu.vn

Received August 2024; accepted November 2024

Abstract. *A translation system, regardless of the approach, needs a quality bilingual corpus to train the translation model. Quality bilingual corpora for rich resource language pairs are currently available to serve the research community. However, for low resource language pairs such as Chinese-Vietnamese, quality bilingual resources for research are not many, and the quality of the corpus is not high. In this paper, we propose to initialize a Chinese-Vietnamese bilingual corpus from the famous subtitle website TED-Talks. We use this corpus for neural machine translation experiments, and the initial results show that the quality of the corpus is satisfactory, higher than that of the TED_MULTI_TRANSLATE corpus on the same translation system.*
Keywords: Bilingual corpus, Chinese-Vietnamese machine translation, Subtitles, TED-Talks

1. **Introduction.** Bilingual corpus is essential for many problems in the field of natural language processing in general and machine translation in particular. To have a machine translation system that allows users to input a source sentence and the system returns a target sentence, such as Google Translate or GPT, requires that the system has a quality translation model. To have this translation model, machine translation systems such as SMT or NMT need a large and quality bilingual corpus to train the translation model. Currently, for rich resource language pairs such as English-Chinese, English-German, large bilingual corpora are available to serve the machine translation research community. There are some bilingual corpora of rich resource language pairs such as English-Chinese UM-Corpus [1], Chinese-English [2], English-German DiscoGeM 2.0 [3]. However, for low resource language pairs such as Chinese-Vietnamese, large corpora to serve research are not many, and the quality of the corpora is limited. Therefore, it is extremely necessary to create a bilingual corpus for low resource language pairs in general and for the Chinese-Vietnamese language pair in particular.

There are many methods to build a bilingual corpus, depending on the purpose of using the corpus, the implementation method is different, it can be fully automatic, semi-automatic (combining automatic with manual) or fully manual. In this paper, we propose to build a fully automatic bilingual corpus, and the purpose of initializing the bilingual corpus is to serve the machine translation research community, specifically Chinese-Vietnamese machine translation.

DOI: 10.24507/icicelb.16.07.727

Selecting resources to exploit bilingual corpus is also an important issue. A poor data source will lead to a poor quality of the collected corpus. Currently, data sources from the web are often chosen for exploitation because of many aspects, such as a large number of articles, and constantly updated content. For Chinese-Vietnamese bilingual websites, there are currently some popular major news sites such as https://www.vietnamplus.vn/, https://baobinhduong.vn/, and https://dangcongsan.vn/. However, through experimental surveys, we found that these news sites are not good resources for automatic bilingual corpus exploitation because bilingual articles are not really parallel, the translation is freestyle, and Chinese articles tend to be more concise than Vietnamese articles.

In this paper, we choose a resource to exploit with better quality and relatively richer content, which is TED-Talks. The Chinese-Vietnamese bilingual corpus of TED-Talks is richer and more parallel than the websites mentioned above. The main problems to be solved when exploiting bilingualism from subtitles websites are 1) characteristics of subtitles websites, and 2) differences in dialogue segments between source and target languages. These are the two main problems that we focus on solving in this paper.

The bilingual corpus will be used to experiment with neural machine translation. Also, we contribute a part of this corpus to the Chinese-Vietnamese machine translation research community. In addition, the experimental results are compared with the bilingual corpus exploited from TED, which is TED_MULTI_TRANSLATE [4]. The structure of the paper will be presented as follows. Section 2 presents the basic knowledge related to the problem of building Chinese-Vietnamese bilingual corpus, constructing corpus from subtitles websites. The method of building bilingual corpus and experimental results are presented in Sections 3 and 4. The conclusion will be presented in Section 5.

2. Background Knowledge.

2.1. Bilingual corpus construction from movie subtitles.

The two most recent representative works on bilingual data generation from subtitles are [5] and [6].

Jafari [5] has generated many bilingual datasets from multilingual movie subtitle websites. One of the biggest contributions of this work is to solve the problem of movie subtitle synchronization. When building a bilingual corpus for a language pair, we usually download the source language and the target language, then align the documents and align the sentences. However, since this work builds a dataset for many language pairs, it will take a long time to download the documents first and then perform the alignment. The author proposed to verify the subtitles before downloading. If the two subtitles are not synchronized, another subtitle of the video will be processed until a matching subtitle is found, and then the subtitle will be downloaded.

Zeroual and Lakhouaja [6] built the MulTed multilingual dataset with more than 100 languages, including Chinese and Vietnamese. The data source is the NetFlix subtitle page. The multilingual dataset is aligned at the sentence level with English as the axis language, and the sentence alignment is performed between English and other languages. Thus, this corpus already exists in two corpora related to Chinese and Vietnamese, namely English-Chinese and English-Vietnamese. However, this corpus has not been published yet.

2.2. Chinese-Vietnamese bilingual corpus construction.

Zhao et al. [7] used a Sino-Vietnamese dictionary and a Vietnamese-Chinese phrase table to automatically translate Vietnamese into Chinese. Their machine translation corpus was mainly a Sino-Vietnamese dictionary and a small Vietnamese-Chinese bilingual corpus to train the Vietnamese-Chinese phrase table. However, these data were not published.

Next is the work of Tran et al. [8], who used OpenSubtitles 2016 as a source of data for mining Chinese-Vietnamese bilingual corpus. However, this corpus still has many errors such as mismatched sentences, translation errors, free translation, and font errors. We

also surveyed the data source from this website, but the quality is not really good; there are quite a few misalignment errors. Our work also mines movie subtitle data, but it is taken directly from the TED-Talks movie website, not from intermediate sites like [8].

The most typical example in the field of building a Chinese-Vietnamese bilingual corpus is Tran et al. [9-11]. In [9], the authors built a semi-automatic Chinese-Vietnamese corpus, the data source was taken from Chinese conversation books and Chinese learning forums, and the corpus has a conversational style, relatively short sentence length, and quite high translation quality up to 35 BLEU points. In [10], the authors used the data source from the Glosbe multilingual dictionary. In this corpus, after each vocabulary and its meaning, there will be bilingual examples related to that vocabulary. The examples are taken from many sources, such as subtitles websites and especially the Bible. Most recently, in [11], the authors used the NetFlix subtitles website as a data source for bilingual exploitation. In this paper, we also use subtitles websites to exploit, specifically TED-Talks. Due to the different characteristics of these two subtitles websites, our data construction method is also different from [11].

2.3. **TED-Talks Open Translation project.** TED Conferences LLC is an American media company that posts talks online with the slogan "ideas worth spreading". Their main product is a show where the organization invites big thinkers and doers to talk about their life. The topics from TED-Talks are usually original researchers, social science, and philosophy. The language shared in the show is English, and there is a project that translates the subtitles from the video into different languages. To ensure the quality of the transcript translation, Figure 1 shows the process where a source transcript (usually English) gets translated by the project's contributor into the destination language; the translation needs to be carefully reviewed before publishing.

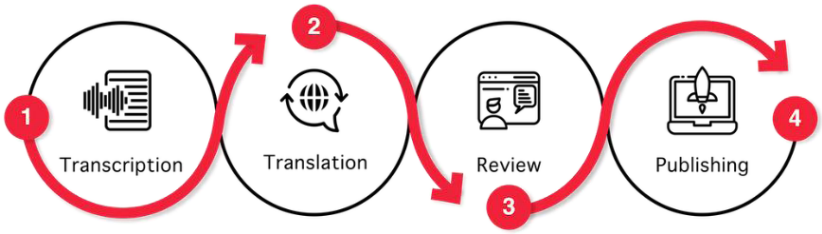

FIGURE 1. TED-Talks Open Translation process from transcription to publishing, official visualization from the project's webpage

3. **Constructing Bilingual Corpus.** Figure 2 presents the general process of constructing bilingual corpus from subtitles.

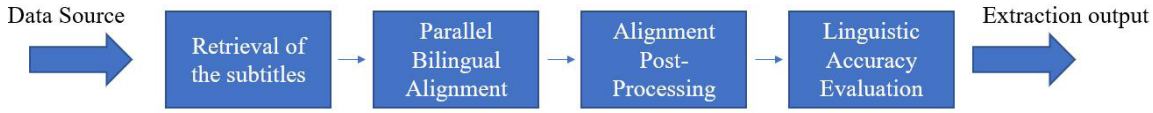

FIGURE 2. Basic steps to build a bilingual database used in this research

3.1. **Retrieval subtitles from TED-Talks.** There are already some datasets that use this source, including the Chinese-Vietnamese language pair, and TED_HRLR_Translate from Tensorflow is one of them, but the subtitles from this dataset are up until 2018. There are plenty of videos that have been released since then. We decided to crawl and

build the bilingual parallel data set with our process to maximize the amount of data we can get from this.

TED stores its subtitles in JSON format rather than a standard subtitle format. Each JSON entry is organized into paragraphs, with each paragraph containing cues that include both the text and its associated timeframe. We will use these timeframes to synchronize the text with its translation. Figure 3 shows an example of TED's data structure.


```
"cues": [
    {
        "time":7954,
        "text":"几千年前，罗马人发明了一种\n得以让他们建设庞大文明的材料。"
    },
    {
        "time":15329,
        "text":"老普林尼夸赞一个由这种材料搭建的海堤"
    },
    {
        "time":19413,
        "text":"称之为“坚不可摧，并且日益坚固”。"
    },
```


FIGURE 3. An example of data structure of a TED-Talks

3.2. **Parallel bilingual alignment.** Each JSON file contains an array of objects called paragraphs, with each paragraph containing an array of sub-objects known as cues. The TED-Talks Transcript features several considerations for the extraction process.

- The number of paragraphs and cues within each paragraph can vary between JSON files.
- The splitting of long sentences into shorter entries occurs not only in Vietnamese but also in Chinese. This separation does not adhere to specific rules and primarily depends on the editor's experience and preferences. However, punctuation is retained in the entries, facilitating the assembly of complete sentences after extraction.
- The entries may also contain special characters, the most notable being "¥n" (the newline escape character), which must be removed during file reading.

Raw data after being collected will be saved in JSON format. These files are read in for processing. Storing data in paragraphs is not beneficial if we only want to extract bilingual data from temporal data. We are only interested in entries so the paragraph will be ignored. Table 1 and Table 2 illustrate the charateristics.

TABLE 1. Extracted information from the Vietnamese transcript file from the video "The material that could change the world. . .for a third time"

Time	Content
34454	Ngày nay, những con đường, vỉa hè, những cây cầu, và các tòa nhà chọc trời
38412	được làm từ vật liệu tương tự, dù ít bền hơn. Đó là bê tông.
43412	Có ba tấn bê tông cho mỗi người trên Trái Đất.

TABLE 2. Extracted information from the Chinese transcript file from the video "The material that could change the world...for a third time"

Time	Content
34454	今天的道路、人行道、桥和摩天大楼
34454	由类似、没有那么耐用的材料所制，
43412	世上每个人有3 吨混凝土。

3.3. **Alignment post-processing.** The current synchronized parallel text is still dirty and needs to be refined. The purposes of this step are

- To merge incomplete entries to be a complete sentence: We used a set of rules based on the observed data, from those rules, apply to merge or ignore the data.
- To pre-normalize data: We remove unwanted Unicode characters, apply padding for punctuation, strip the redundant spaces, etc.
- To split entries into smaller sentences and de-duplicate: We remove these misaligned entries by counting the number of sentences and the question type of source language and target language.
- Chinese punctuation uses a different set of punctuation marks from European languages and has shapes that are derived from both Western and Chinese sources (Table 3).

TABLE 3. Some of the punctuation used in Chinese and its Latin equivalency

Punctuation mark	Punctuation in Chinese	Punctuation in Latin-based languages
Full stop	。	.
Quotation mask	⌈...⌋, ¬...⌙ , "..."	" ", ' '
Enumeration comma	、	,
Question mark	?	?
Exclamation mark	!	!

3.4. **Linguistic accuracy evaluation.** The aligned corpus is still substantial, necessitating the development of a strategy for evaluation. We propose the following solution.

- Utilize the Google Translate API to translate the Chinese data.
- Calculate the BLEU score for each translation result against the Vietnamese side of the corpus. If the BLEU score exceeds a threshold – specifically, 0.18 – we can consider the alignment to be correct.

We selected 0.18 as our threshold because it corresponds to the highest density in the distribution of BLEU scores across sentences, providing a reliable measure for alignment quality.

3.5. **Extraction output.** After filtering the corpora, Figure 4 represents the sentence length distribution of the desired dataset and Table 4 shows the number of final sentence pairs.

Regarding the sentence length distribution, dialogue in movies often features short sentences to enhance pacing and maintain audience engagement, as concise lines are easier to follow in fast-paced scenes. This brevity mirrors real-life speech, making interactions feel more authentic and relatable.

Looking at the percentage after validation, we see that only around 70% of the pair are accepted, it is because the translation quality between subtitles created by humans and those generated by tools like Google Translate often differs due to several factors. Human

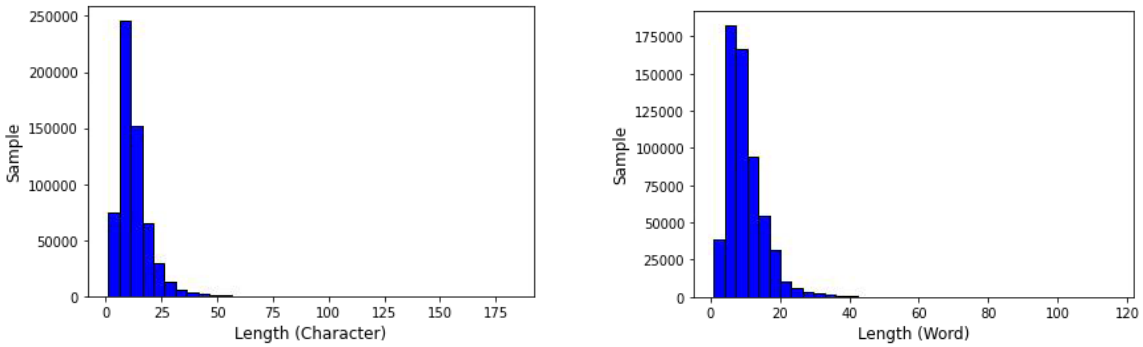

FIGURE 4. Sentence length distribution from Chinese and Vietnamese on TED-Talks

TABLE 4. Extraction output from TED-Talks Transcript source

Corpora	Before validation	After validation	Percentage
TED-Talks	161.662	113.012	69.90%

translators consider context, cultural nuances, idiomatic expressions, and emotional tone, which helps them convey meaning more accurately and effectively, even in short sentences. In contrast, automated tools typically focus on literal translations, often struggling with subtleties and context, leading to awkward or inaccurate results. This can result in lower BLEU scores, as human translations are better at maintaining the intended message and flow, especially when sentence length and conversational tone are taken into account. However, to achieve overall performance, we accept this metric of validation.

4. Experiments.

4.1. **Corpora.** We perform separate training for two datasets. For comparison purposes, we also perform training for one previously built dataset TED_MULTI_TRANSLATE for comparison with the TED-Talks built by our proposed process.

Each corpus is divided into three subsets: training set, testing set, and validation set. The method of division is as follows: Every 25 sentences, the 24th sentence is divided into the testing set, the 25th sentence is divided into the training set, and the remaining sentences are used as the training set. Table 5 lists the number of sentence pairs for each subset of each corpus.

TABLE 5. Number of sentence pairs for each source and each subset

Corpora	Number of pairs	Training size	Testing size	Validation size
TED-Talks	113,188	103,569	4,707	4,912
TED_MULTI_TRANSLATE	159,588	146,025	6,637	6,926

4.2. **Toolkits used in experiments.** For experiment purposes, we used Fairseq [12] to train and validate machine translation models with our datasets. Fairseq is a publicly available toolkit for sequence modeling, enabling researchers and developers to train personalized models tailored for tasks such as translation, summarization, language modeling, and various other text generation endeavors.

4.3. **NMT models.** In this experimentation, we used two models. The first one is the Fully Convolutional Network proposed by Gehring et al. [13]. The second model is Transformer proposed by Vaswani et al. [14]. The Fully Convolutional Network is chosen for its ability to efficiently capture local patterns in data through convolutional layers, making it well-suited for tasks requiring detailed feature extraction. In contrast, the Transformer model is selected for its powerful attention mechanism, which enables it to effectively process long-range dependencies and contextual information in sequences.

4.4. **Training configuration.** We ran our experiments on Google Colab Pro. The Pro version supports higher RAM capacity and longer training time. We trained our models on a Tesla P100-PCIE-16GB. Table 6 describes the time and number of epochs we used to train.

Table 6. Training time and the number of epochs on Fully Convolutional Network (FCN) and Transformer

Model	TED_MULTI_TRANSLATE	TED (our process)
FCN		
# of epochs	40	40
Training time	∼7 hours	∼5 hours
Transformer		
# of epochs	40	40
Training time	∼3 hours	∼2 hours

4.5. **Experimental results.** Table 7 shows experimental results.

Table 7. BLEU scores

Corpus	FCN	Transformer
TED-Talks Translation	**17.90**	15.98
TED_MULTI_TRANSLATE	12.11	4.43

4.6. **Analysis.** From Table 7, it is easy to see that with the proposed method of extracting data from subtitles in this paper, the BLEU score is much higher than that of the proposed extraction method of the two datasets. The highest BLEU score belongs to the TED-Talks Translation dataset of 17.90 trained on the FCN model. Overall, the model gives good translation results on short or medium-length sentences but not very well on long sentences or sentences containing rare words. Factors that lead to these problems include

- Complicated sentences with duplication structure and repeating of words, makes the model hard to recognize the pattern.
- English words are translated literally, i.e., "marathon" translated to be "chạy" without using the right word for the sport "điền kinh".

5. **Conclusions.** In this paper, we utilized the Chinese-Vietnamese bilingual data sourced from the TED-Talks subtitles website to support the Chinese-Vietnamese machine translation research community. The method for constructing the bilingual corpus is not only applicable to the Chinese-Vietnamese language pair but can also be extended to many other language pairs supported by TED, particularly those of similar types, such as Chinese and Vietnamese. This corpus has been employed to experiment with neural machine translation, yielding an initial BLEU score of approximately 18 points. While this result surpasses that of the TED_MULTI_TRANSLATE corpus, it remains relatively low compared to general machine translation quality.

In the near future, we aim to enhance the quality of the TED corpus and continue exploring additional sources of Chinese-Vietnamese bilingual data. Our goal is to enrich the bilingual corpus landscape, thereby contributing to the advancement of machine translation research.

REFERENCES

[1] L. Tian, D. F. Wong, L. S. Chao, P. Quaresma, F. Oliveira, S. Li, Y. Wang and Y. Lu, UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation, *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014.

[2] S. Jie, Construction of large-scale Chinese-English bilingual corpus and sentence alignment, in *Application of Big Data, Blockchain, and Internet of Things for Education Informatization. BigIoT-EDU 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, M. A. Jan and F. Khan (eds.), vol.466, Springer, Cham, DOI: 10.1007/978-3-031-23947-2_42, 2023.

[3] F. Yung, M. Scholman, Š. Zikánová and V. Demberg, DiscoGeM 2.0: A parallel corpus of English, German, French and Czech implicit discourse relations, *Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp.4940-4956, 2024.

[4] Y. Qi, S. Devendra, F. Matthieu, P. Sarguna and N. Graham, When and why are pre-trained word embeddings useful for neural machine translation, *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, LA, USA, pp.529-535, 2018.

[5] F. Jafari, Generating multilingual parallel corpus using subtitles, *arXiv.org*, arXiv: 1804.03923, 2018.

[6] I. Zeroual and A. Lakhouaja, MulTed: A multilingual aligned and tagged parallel corpus, *Applied Computing and Informatics*, vol.18, no.1/2, pp.61-73, DOI: 10.1016/j.aci.2018.12.003, 2022.

[7] H. Zhao, T. Yin and J. Zhang, Vietnamese to Chinese machine translation via Chinese character as pivot, *Proc. of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, Taipei, Taiwan, pp.250-259, 2013.

[8] H.-A. Tran, Y. Guo, P. Jian, S. Shi and H. Huang, Improving parallel corpus quality for Chinese-Vietnamese statistical machine translation, *Journal of Beijing Institute of Technology*, vol.27, no.1, pp.127-136, 2018.

[9] P. Tran, D. Dinh and L. H. B. Nguyen, Word re-segmentation in Chinese-Vietnamese machine translation, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol.16, no.2, pp.1-22, DOI: 10.1145/2988237, 2016.

[10] P. Tran, T. Nguyen, D.-H. Vu, H.-A. Tran and B. Vo, A method of Chinese-Vietnamese bilingual corpus construction for machine translation, *IEEE Access*, vol.10, pp.78928-78938, DOI: 10.1109/ACCESS.2022.3186978, 2022.

[11] P.-N. Nguyen and P. Tran, Constructing a Chinese-Vietnamese bilingual corpus from subtitle websites, *International Journal of Intelligent Information and Database Systems*, vol.16, no.4, pp.385-408, DOI: 10.1504/IJIIDS.2024.10065439, 2024.

[12] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier and M. Auli, Fairseq: A fast, extensible toolkit for sequence modeling, *arXiv.org*, arXiv: 1904.01038, 2019.

[13] J. Gehring, M. Auli, D. Grangier, D. Yarats and Y. N. Dauphin, Convolutional sequence to sequence learning, *Proc. of the 34th International Conference on Machine Learning (ICML'17)*, Sydney, NSW, Australia, pp.1243-1252, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, *Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, pp.6000-6010, 2017.