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, thviet@uneti.edu.vn, anhuni1006@gmail.com

INTRODUCTION

- Vietnamese is known as a low-resource language for multilingual parallel corpus. The multilingual parallel corpus for Vietnamese is very rare, so the quality of machine translation for Vietnamese is not good.
- The high-quality multilingual parallel corpus is an important key for improving the quality of the machine translation system.
- A good quality machine translation system that translates text from Vietnamese into Chinese, Laos, or Khmer is more essential than ever in order to support the information exchange of social political organizations, economic groups, and people between Vietnam and China, Lao, Cambodia.

⇨This has motivated us to focus on building a high-quality multilingual corpus, including Vietnamese, Chinese, Laos, and Khmer languages and publish for the research community

Building Multilingual Parallel Corpus

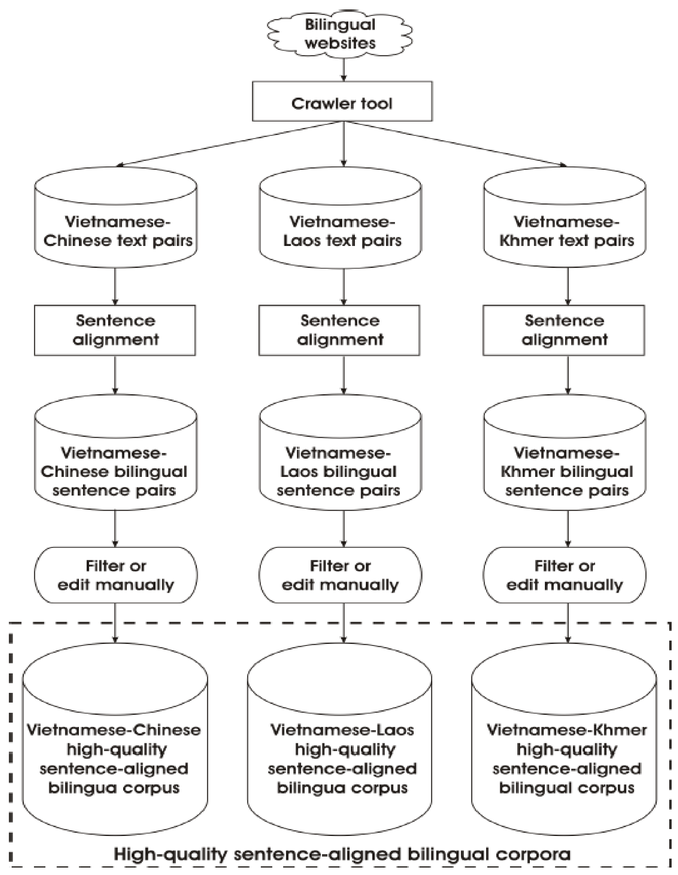


Figure 1: The method of constructing high-quality sentence-aligned multilingual parallel corpus.

Collecting Parallel Text Pairs

- We crawl bilingual websites in the news domain for Vietnamese-Chinese, Vietnamese-Laos, and Vietnamese-Khmer language pairs. They are stored by posted month.

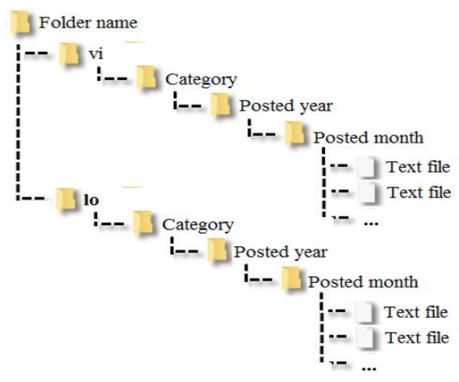
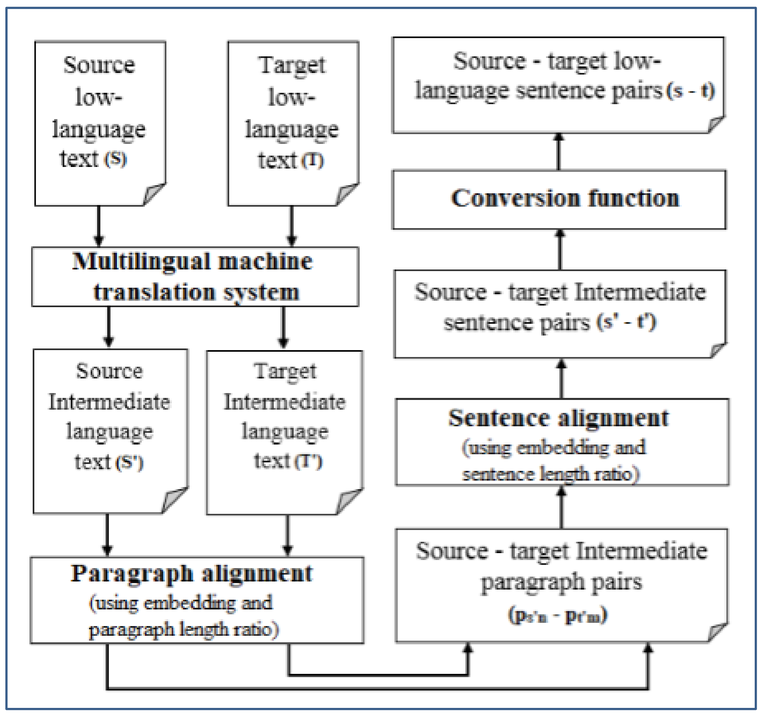


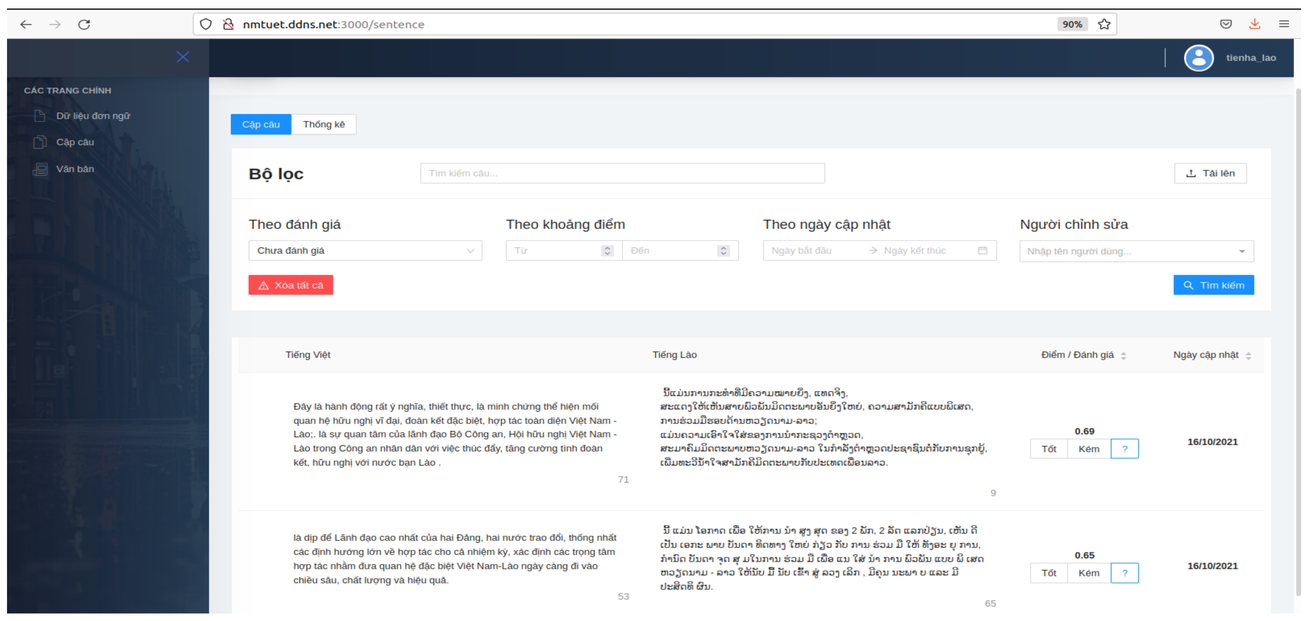
Figure 2: The text stored tree for Vietnamese - Laos pair

Sentence Alignment



Manual Data Reviewing

- We built the online tool[1] for manually reviewing automatic aligned bilingual sentences.
- Each bilingual sentence pair is assigned with one of two levels by reviewers:
  - Good level: This level is chosen if two sentences in the pair are the translation of each other. For bilingual sentence pairs that are easy to modify to get good pairs, annotators will select good level after modifying them.
  - Bad level: This level is chosen if they are not translation of each other or difficult to modify to get good pairs.



Experiments

Parallel Corpus Statistics

Three parallel corpus have been built: Vietnamese-Laos (Lo-Vi), Vietnamese-Khmer (Kh-Vi), Vietnamese-Chinese (Zh-Vi). 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

In Table 1: *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.

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

In Table 2: *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.

Intrinsic Evaluation Results

In Table 4: *#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.

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

Extrinsic Evaluations

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

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 (0.15%)	257 (12.84%)	770 (38.46%)	627 (31.32%)	223 (11.14%)	122 (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 (0.00%)	427 (21.33%)	739 (36.91%)	530 (26.47%)	186 (9.29%)	120 (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 (0.05%)	501 (25.02%)	865 (43.21%)	357 (17.83%)	159 (7.94%)	119 (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	Δ(%)
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

Conclusion and Future Work

In this work, 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.