Interspeech 2022
18-22 September 2022, Incheon, Korea

A Vietnamese-English Neural Machine Translation System

Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, Dat Quoc Nguyen

VinAI Research, Vietnam

{v.thiennh7, v.duynht1, v.duypv1, v.duyntc1, v.hieutm4,
v.manhlt3, v.tinvd12, v.hungbh1, v.dinhpq2, v.datnq9}@vinai.io

Abstract

We present VinAI Translate—a system that integrates state-of-the-art deep learning technologies in speech and natural language processing to translate speech and text between Vietnamese and English. Experimental results show that our system obtains a state-of-the-art performance for each translation direction, and outperforms Google Translate in both automatic and human evaluations.

Index Terms: Machine translation; Automatic speech recognition; Text-to-speech; Vietnamese; English.

1. Introduction

The last two decades have witnessed rapid economic growth in Vietnam which is now an attractive destination for trade and investment. Due to the language barrier, foreigners generally use automatic machine translation systems to translate Vietnamese speeches and texts into their native language or another language they are familiar with—usually the global language English—so they could quickly catch up with ongoing events in Vietnam. Thus the demand for high-quality Vietnamese-English speech and text translation has rapidly increased [1].

Recently, the performance of neural models in Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-To-Speech (TTS) has been pushed to a cutting-edge boundary, which is relatively competitive with the human level. Our goal is to employ modern ASR, MT and TTS approaches to build an application that helps translate speech and text from Vietnamese to English and vice versa at a high-level quality.

In this paper, we introduce VinAI Translate—a neural machine translation system that translates speech and text between Vietnamese and English. Our system is a user-friendly and interactive interface engine that integrates advanced deep learning models for ASR, MT and TTS. Experiments show that our system obtains state-of-the-art performance results in both automatic and human evaluations. Our system is available at: `https://vinai-translate.vinai.io`.

2. Our VinAI Translate system

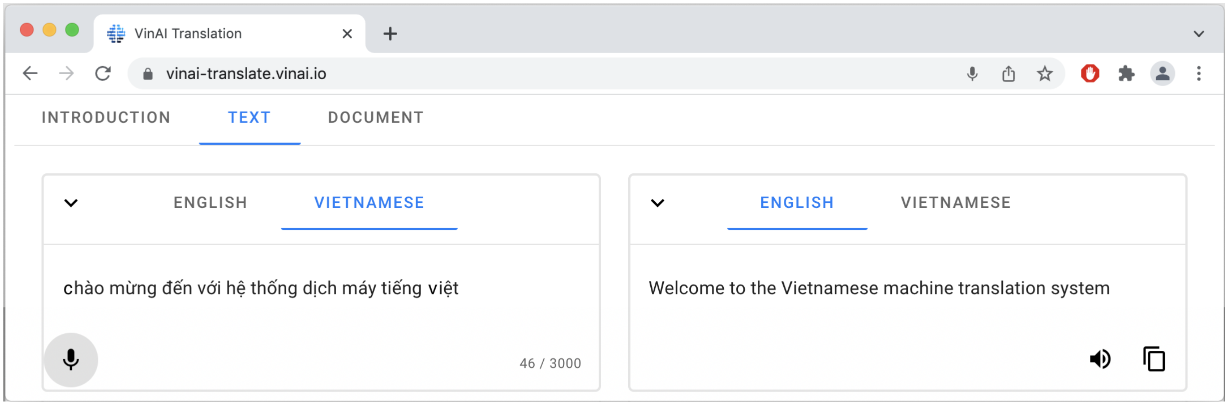
Our system consists of four components, including User Interface (UI) and three backend components of ASR, MT and TTS. The UI component takes a speech or text in the source language from a user as input. If the input is a speech, UI passes it to the ASR component. Then ASR converts the input speech to text. The MT component translates either the input text taken from UI or the text output from ASR into a text in the target language. The translated text output from MT is then fed as input into the TTS component that generates a synthesized speech. Finally, UI responds to the user by displaying the translated text output from MT as well as the synthesized speech output from TTS.

The UI component: Figure 1 shows our system's UI. To input a speech, the users need to press the microphone icon at the bottom left of the input box; while to input a text, the users need to type or copy and paste an existing text into the input box. The backend is then activated to process the input. UI then displays the backend's translated text in the output box. The users need to press the speaker icon at the bottom right of the output box to hear the synthesized speech of the translated text. For example, a user could input a Vietnamese speech of "chào mừng đến với hệ thống dịch máy tiếng việt", and our system produces an English translation output of "Welcome to the Vietnamese machine translation system" in both forms of text and speech.

The ASR component: To develop our ASR component, for Vietnamese, we train Conformer-CTC [2] using our in-house 5700-hour dataset augmented by noise injection and intensity adjustment approaches, and we obtain the word error rate (WER) at about 1.4% on our internal test set. For English, we train Conformer-CTC on the Librispeech training set [3] and obtain WER at 1.8% on the Librispeech test-clean set. For inference in each language, we further incorporate a 6-gram Byte-Pair-Encoding-based language model [4] into the decoder to enhance the ASR performance.

The MT component: We first fine-tune the pre-trained sequence-to-sequence model mBART [5] using 3M training sentence pairs from the high-quality dataset PhoMT [1] for English-to-Vietnamese translation. From each English-Vietnamese sentence pair in "noisy" datasets CCAligned [6] and WikiMatrix [7], we employ the fine-tuned model to translate the English sentence into Vietnamese. We only select from CCAligned and WikiMatrix the English-Vietnamese sentence pairs that have a BLEU score [8] between the Vietnamese-translated variant of the English source sentence and the Vietnamese target sentence in the range of 0.15 to 0.95, resulting in 6M pairs. We thus have a collection of 3M + 6M = 9M "high-quality" sentence pairs. To simulate the ASR output, for each translation direction English-to-Vietnamese (EN-VI) or Vietnamese-to-English (VI-EN), we also augment the collection with additional 9M sentence pairs where we lowercase and remove punctuations from the source sentences while keeping the target sentences intact. We fine-tune mBART for each translation direction using 9M + 9M = 18M sentence pairs to develop our MT component.

The TTS component: The TTS component first converts the translated text into phonemes based on their pronunciation and text normalization rules. For Vietnamese, our TTS employs Glow-TTS [9] to predict mel-spectrogram from input phonemes. Here, we modify the input of the Glow-TTS model to be well-fitted with Vietnamese by using the Vietnamese phoneme dictionary. For English, our TTS employs Tacotron2 [10] to predict mel-spectrogram from input phonemes. Our TTS then uses HiFi-GAN [11] to generate efficient and high-fidelity speech synthesis from the predicted mel-spectrogram.



chào mừng đến với hệ thống dịch máy tiếng việt

Welcome to the Vietnamese machine translation system

Figure 1: *Our system's UI on web interface.*

Table 1: *BLEU scores.*

Model	Validation set		Test set	
	EN-VI	VI-EN	EN-VI	VI-EN
Google Translate	40.10	36.89	39.86	35.76
PhoMT [1]	44.32	40.88	43.46	39.78
VinAI Translate	**45.31**	**41.41**	**44.29**	**40.42**

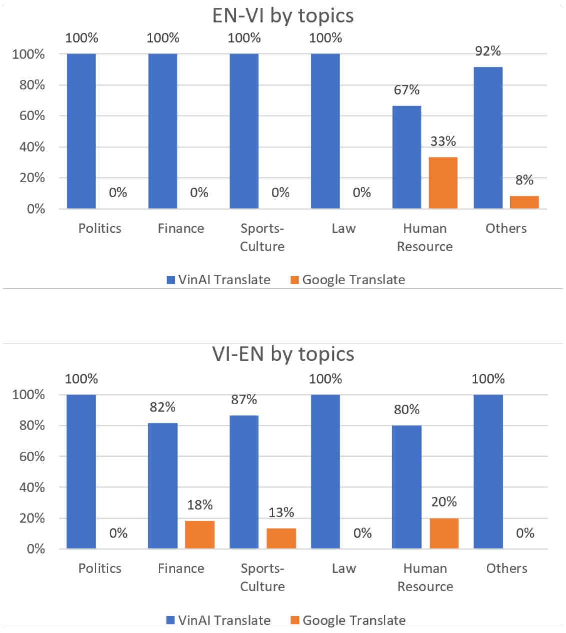


Figure 2: *Human evaluation results.*

3. Evaluation

Automatic evaluation results: We evaluate our VinAI Translate system on the benchmark PhoMT's validation and test sets [1]. Table 1 presents BLEU scores obtained by our system and Google Translate as well as the highest BLEU scores reported in the PhoMT paper. Table 1 shows that our system reaches a new state-of-the-art performance for each translation direction, outperforming Google Translate by a large margin.

Human evaluation results: For each translation direction, we sample 100 paragraphs from different domains, translate them using Google Translate and our VinAI Translate and anonymously shuffle the translation outputs. We ask 11 external annotators to choose which translation they think is better. Figure 2 shows the final results, where VinAI Translate consistently outperforms Google Translate for both translation directions.

4. Conclusions and future work

In this paper, we have introduced VinAI Translate to translate speech and text between Vietnamese and English. Automatic and human evaluation results show that our system obtains state-of-the-art performances. Future work includes improving the translation performance as well as extending the system to other Vietnamese-centric language pairs. We also publicly release our pre-trained translation models at `https://github.com/VinAIResearch/VinAI_Translate`.

5. References

[1] L. Doan, L. T. Nguyen, N. L. Tran, T. Hoang, and D. Q. Nguyen, "PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation," in *EMNLP*, 2021.

[2] P. Guo *et al.*, "Recent Developments on Espnet Toolkit Boosted By Conformer," in *ICASSP*, 2021.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[4] W. Hu, Y. Luo, J. Meng, Z. Qian, and Q. Huo, "A study of bpe-based language modeling for open vocabulary latin language ocr," in *ICFHR*, 2020.

[5] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of ACL*, vol. 8, 2020.

[6] A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn, "CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs," in *EMNLP*, 2020.

[7] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia," in *EACL*, 2021.

[8] M. Post, "A Call for Clarity in Reporting BLEU Scores," in *WMT*, 2018.

[9] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," in *NeurIPS*, 2020.

[10] J. Shen *et al.*, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *ICASSP*, 2018.

[11] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *NeurIPS*, 2020.