Recent Advances in English-Vietnamese Text and Speech Translation

Dat Quoc Nguyen
Head of NLP, VinAI
https://datquocnguyen.github.io
Motivation

- The demand for high-quality Vietnamese-English machine translation has rapidly increased
  - Strategy: Employ modern ASR, MT and TTS approaches to build an application that helps translate text and speech between Vietnamese and English at a high-level quality
Motivation

• Issues:
  • High-quality Vietnamese-English parallel corpora are either not publicly available or small-scale
  • Larger Vietnamese-English parallel corpora are noisy

• Approach:
  • Construct high-quality and large-scale parallel datasets
    • PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation (EMNLP 2021) [1]
    • PhoST: A High-Quality and Large-Scale Dataset for English-Vietnamese Speech Translation (InterSpeech 2022) [2]
  • Train state-of-the-art sequence-to-sequence models
    • VinAI Translate: A Vietnamese-English Neural Machine Translation System (InterSpeech 2022: Show & Tell) [3]
Outline

• VinAI Translate: A Vietnamese-English Neural Machine Translation System
  • PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation
• PhoST: A High-Quality and Large-Scale Dataset for English-Vietnamese Speech Translation
Introduction

Welcome to the Vietnamese machine translation system.
Automatic Speech Recognition

- For Vietnamese
  - Obtain the word error rate (WER) at about 1.4% on an internal test set
- For English
  - Train Conformer-CTC on the Librispeech training set [5] and obtain WER at 1.8% on the Librispeech test-clean set
- For inference in each language: Incorporate a 6-gram Byte-Pair-Encoding-based language model [6] into the decoder to enhance the ASR performance
Text-to-Speech

• Convert the translated text into phonemes based on their pronunciation and text normalization rules
• Predict mel-spectrogram from input phonemes
  • Employ & modify Glow-TTS [7] for Vietnamese, using a Vietnamese phoneme dictionary
  • Employ Tacotron2 [8] for English
• Use HiFi-GAN [9] to generate efficient and high-fidelity speech synthesis from the predicted mel-spectrogram
Machine Translation

- Approach: Fine-tune the pre-trained Seq2Seq model mBART [10] on a large-scale parallel dataset
- Construct PhoMT—a high-quality and large-scale Vietnamese-English parallel dataset
  1. Collecting parallel document pairs
  2. Pre-processing
  3. Aligning parallel sentence pairs
  4. Post-processing
Machine Translation

- PhoMT: Collecting parallel document pairs
Machine Translation

• PhoMT: Pre-processing
  • Manually inspect and remove low-quality document pairs from OpenSubtitles domain
  • Filter English paragraphs inside Vietnamese documents (and vice versa)
  • Perform sentence segmentation using VnCoreNLP [11] and Stanford CoreNLP [12]
• PhoMT: Aligning parallel sentence pairs
  • Translate English source sentences into Vietnamese using Google Translate
  • Align between translated source sentences and target sentences using 3 toolkits: Hunalign [13], Gargantua [14], Bleualign [15]
  • Select pairs that are aligned by at least 2/3 toolkits
Machine Translation

- PhoMT: Post-processing
  - Split the dataset into train/validation/test sets
  - Manually inspect validation and test sets and remove misaligned and low-quality sentence pairs (0.8%)
- PhoMT: A high-quality and large-scale Vietnamese-English parallel dataset consisting of 3.02M pairs

<table>
<thead>
<tr>
<th>Domain</th>
<th>Total</th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#doc</td>
<td>#pair</td>
<td>#pair</td>
<td>#en/s</td>
</tr>
<tr>
<td>News</td>
<td>2559</td>
<td>41504</td>
<td>40990</td>
<td>24.4</td>
</tr>
<tr>
<td>Blogspot</td>
<td>1071</td>
<td>93956</td>
<td>92545</td>
<td>25.0</td>
</tr>
<tr>
<td>TED-Talks</td>
<td>3123</td>
<td>320802</td>
<td>316808</td>
<td>19.8</td>
</tr>
<tr>
<td>MediaWiki</td>
<td>38969</td>
<td>496799</td>
<td>490505</td>
<td>26.0</td>
</tr>
<tr>
<td>WikiHow</td>
<td>6616</td>
<td>513837</td>
<td>507379</td>
<td>18.9</td>
</tr>
<tr>
<td>OpenSub</td>
<td>3312</td>
<td>1548971</td>
<td>1529772</td>
<td>9.7</td>
</tr>
<tr>
<td>All</td>
<td>55650</td>
<td>3015869</td>
<td>2977999</td>
<td>15.7</td>
</tr>
</tbody>
</table>

https://github.com/VinAIResearch/PhoMT
Machine Translation

- Fine-tune mBART on the PhoMT training set of ~3M pairs for English-to-Vietnamese
- From each English-Vietnamese sentence pair in “noisy” datasets CCAigned [16] and WikiMatrix [17]
  - Employ the fine-tuned model to translate the English sentence into Vietnamese
  - Select pairs with a BLEU score between the Vietnamese-translated variant and the Vietnamese target sentence ranging from 0.15 to 0.95, resulting in 6M pairs
- A collection of 3M + 6M = 9M “high-quality” sentence pairs
- Simulate the ASR output: Lowercase and remove punctuations from the source sentences while keeping the target sentences intact, obtaining 9M pairs for each translation direction
- For each translation direction: 9M + 9M = 18M sentence pairs
Machine Translation

- Fine-tune mBART for each translation direction using 18M sentence pairs
  - Reduce mBART vocabulary from 250K tokens to 90K tokens belonging to English and Vietnamese
- Publicly released: [https://github.com/VinAIResearch/VinAI_Translate](https://github.com/VinAIResearch/VinAI_Translate)

<table>
<thead>
<tr>
<th>Model</th>
<th>#params</th>
<th>Max length</th>
</tr>
</thead>
<tbody>
<tr>
<td>vinai/vinai-translate-vi2en</td>
<td>448M</td>
<td>1024</td>
</tr>
<tr>
<td>vinai/vinai-translate-en2vi</td>
<td>448M</td>
<td>1024</td>
</tr>
</tbody>
</table>

- Pre-trained VinAI Translate models can be used with the popular open-source library transformers
- These pre-trained models are used in the MT component of the VinAI Translate system: [https://vinai-translate.vinai.io](https://vinai-translate.vinai.io)
- Users can also try these models at: [https://huggingface.co/spaces/vinai/VinAI_Translate](https://huggingface.co/spaces/vinai/VinAI_Translate)
PhoMT evaluation results

- Experimental results on the PhoMT validation and test sets while using the PhoMT training set of 2.97M pairs for training

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>En-to-Vi</td>
<td>Vi-to-En</td>
</tr>
<tr>
<td></td>
<td>TER↓ BLEU↑</td>
<td>TER↓ BLEU↑</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Translate</td>
<td>45.86</td>
<td>40.10</td>
</tr>
<tr>
<td>Bing Translator</td>
<td>45.36</td>
<td>40.82</td>
</tr>
<tr>
<td>Transformer-base</td>
<td>42.77</td>
<td>43.01</td>
</tr>
<tr>
<td>Transformer-big</td>
<td>42.13</td>
<td>43.75</td>
</tr>
<tr>
<td>mBART</td>
<td>41.56</td>
<td>44.32</td>
</tr>
</tbody>
</table>

- mBART achieves the best performances, in both translation directions and on all metrics
- Neural MT baselines outperform automatic translation engines
PhoMT evaluation results

- BLEU scores of Transformer-base on the Vi- to-En validation set when varying training sizes on PhoMT

- Sample a set of 1.55M non-duplicate Vietnamese-English sentence pairs from OPUS’s OpenSubtitles, which has the same size as the PhoMT’s OpenSubtitles training subset:
  - OPUS’s OpenSubtitles: 29.72 BLEU
  - PhoMT’s OpenSubtitles: 31.11 BLEU

Our curation effort paid off!
VinAI Translate evaluation results

- **Automatic evaluation results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN-VI</td>
<td>VI-EN</td>
</tr>
<tr>
<td>Google Translate</td>
<td>40.10</td>
<td>36.89</td>
</tr>
<tr>
<td>PhoMT</td>
<td>44.32</td>
<td>40.88</td>
</tr>
<tr>
<td>VinAI Translate</td>
<td>45.31</td>
<td>41.41</td>
</tr>
</tbody>
</table>

- **Human evaluation results**

**VI-EN by topics**

- Politics: 100% (VinAI Translate), 0% (Google Translate)
- Finance: 82% (VinAI Translate), 18% (Google Translate)
- Sports-Culture: 87% (VinAI Translate), 13% (Google Translate)
- Law: 100% (VinAI Translate), 0% (Google Translate)
- Human Resource: 80% (VinAI Translate), 20% (Google Translate)
- Others: 100% (VinAI Translate), 0% (Google Translate)

**EN-VI by topics**

- Politics: 0% (VinAI Translate), 0% (Google Translate)
- Finance: 0% (VinAI Translate), 0% (Google Translate)
- Sports-Culture: 0% (VinAI Translate), 0% (Google Translate)
- Law: 0% (VinAI Translate), 0% (Google Translate)
- Human Resource: 67% (VinAI Translate), 33% (Google Translate)
- Others: 92% (VinAI Translate), 8% (Google Translate)
Takeaways

- PhoMT—A high-quality and large-scale Vietnamese-English parallel dataset: https://github.com/VinAIResearch/PhoMT
Outline

• VinAI Translate: A Vietnamese-English Neural Machine Translation System
  • PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation
• PhoST: A High-Quality and Large-Scale Dataset for English-Vietnamese Speech Translation
Introduction

- No existing research work focuses solely on speech translation to Vietnamese
- Only available 441-hour data of English-Vietnamese speech translation from MuST-C [18] which is a TED-talk-based multilingual dataset
  - 5.63% of the validation set and 4.10% of the test set have an incorrect audio start or end timestamp of an English source sentence
  - 16.15% of the validation set and 9.3% of the test set have misaligned English-Vietnamese sentence pairs
Introduction

• Contributions of this work
  • Present a new high-quality and large-scale English-Vietnamese speech translation dataset, named PhoST, with 508 audio hours
  • Empirically investigate strong neural baselines on PhoST to compare traditional “Cascaded” and modern “End-to-End” approaches
  • Publicly release the PhoST dataset at https://github.com/VinAIResearch/PhoST
PhoST dataset construction

- Dataset construction process includes 5 phases
  1. Collecting audio files and transcripts
  2. Pre-processing and sentence segmentation
  3. Extracting the audio start and end timestamps for each English sentence
  4. Aligning parallel English–Vietnamese sentence pairs
  5. Post-processing
PhoST dataset construction

• Collecting audio files and transcripts
  • Collect audio files and transcripts from the TED2020-v1 corpus [19]
    ⇒ 3120 triplets of (audio file, English transcript document, Vietnamese subtitle document)

• Pre-processing and sentence segmentation
  • Manually check and remove 33 triplets with non-English or displaying songs in audio files
  • Perform sentence segmentation using VnCoreNLP and Stanford CoreNLP for Vietnamese and English documents, respectively
  • Remove all the non-speech artifacts of audience-related information, e.g. "(applause)", "(laugh)" and the like, as well as all the speaker identity from the transcripts
PhoST dataset construction

• Extracting the audio start and end timestamps for each English sentence
  • Employ the Gentle forced aligner [20] to obtain a timestamp for each word token
  • Manually correct the start and end timestamp of 10K English sentences where the Gentle forced aligner cannot detect the timestamp of the first or last word in a sentence

• Aligning parallel English-Vietnamese sentence pairs
  • Translate English source sentences into Vietnamese using Google Translate
  • Align between translated source sentences and target sentences using 3 toolkits: Hunalign, Gargantua, Bleualign
  • Select pairs that are aligned by at least 2/3 toolkits
PhoST dataset construction

- **Post-processing**
  - Split the dataset into train/validation/test sets
  - Manually inspect validation and test sets to remove misaligned between English audio-transcript pairs (0%) and low-quality translation in sentence pairs (0.15%)

- **PhoST dataset statistics**

<table>
<thead>
<tr>
<th>Split</th>
<th>#triplets</th>
<th>#hours</th>
<th>#en/s</th>
<th>#vi/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>327370</td>
<td>501.59</td>
<td>16.55</td>
<td>20.94</td>
</tr>
<tr>
<td>Validation</td>
<td>1933</td>
<td>3.13</td>
<td>17.24</td>
<td>22.22</td>
</tr>
<tr>
<td>Test</td>
<td>1976</td>
<td>3.77</td>
<td>19.23</td>
<td>25.65</td>
</tr>
</tbody>
</table>

- #triplets: the number of triplets
- #hours: the number of audio hours
- #en/s: the average number of word tokens per English sentence
- #vi/s: the average number of syllable tokens per Vietnamese sentence
Speech translation approaches

• Cascaded: English automatic speech recognition (ASR) & English-to-Vietnamese text translation (MT)
  • ASR: Train the Fairseq’s S2T Transformer [21] on the PhoST’s English audio-transcript training set
  • MT: Fine-tune the pretrained sequence-to-sequence model mBART
  • Perform data augmentation to extend the MT training data
    • Convert the trained S2T Transformer’s automatic ASR output into its written form
    • Recover capitalization and punctuation marks
      \[ \Rightarrow 327370 \times 3 = 982110 \] parallel English-Vietnamese sentence pairs

• End-to-end:
  • S2T Transformer
  • The UPC’s speech translation system Adaptor [22] that is the only top performance system at IWSLT 2021 with publicly available implementation
Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casc. (I) mBART w/ our extended dataset</td>
<td>33.65</td>
</tr>
<tr>
<td>(II) mBART w/ PhoMT combination</td>
<td>34.31</td>
</tr>
<tr>
<td>E2E S2T Transformer</td>
<td>29.98</td>
</tr>
<tr>
<td>UPC’s Adaptor</td>
<td>33.30</td>
</tr>
</tbody>
</table>

- (I): mBART fine-tuned on our extended training set
- (II): mBART fine-tuned on a combination of the 3M-pair dataset PhoMT and our extended training set
- Cascaded:
  - The word error rate (WER) computed for the ASR component is 7.06
  - BLEU scores of (I) and (II) computed for the text-to-text MT component with “gold” English source transcript sentences are 36.48 and 37.41, respectively
Experiments

- Compare performances on the MuST-C’s English-Vietnamese training set and PhoST’s training “subset”
  - Same training data size
  - Employ S2T Transformer for ASR
  - Employ the end-to-end Adaptor model for speech translation

<table>
<thead>
<tr>
<th>Training data</th>
<th>WER↓</th>
<th>BLEU↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>MuST-C En-Vi training set</td>
<td>9.09</td>
<td>31.66</td>
</tr>
<tr>
<td>Our sampled training subset</td>
<td>7.44</td>
<td>32.37</td>
</tr>
</tbody>
</table>
Experiments

• An example to demonstrate the qualitative differences between the end-to-end Adaptor models trained on MuST-C and on the PhoST’s training “subset”
  • Input audio of the English sentence: “But on a long wavelength sea, you’d be rolling along, relaxed, low energy.”
  • Output of the end-to-end Adaptor model trained on MuST-C: “Những trên sóng biển dài, bạn sẽ lăn dốc theo, thư giãn, năng lượng thấp.”
  • Output of the end-to-end Adaptor model trained on the subset of the PhoST training set: “Những trên một vùng biển có bước sóng dài, bạn sẽ lăn dốc, thư giãn, ít tổn năng lượng hơn.”
Takeaways

• A high-quality and large-scale dataset with 508 audio hours for English-Vietnamese speech translation
• Compare the “Cascaded” and “End-to-End” approaches using strong baselines
  • “Cascaded” does better than “End-to-End”
• Publicly release the PhoST dataset at: https://github.com/VinAIResearch/PhoST
Thank you!

@VinAI

https://www.vinai.io/
References

1. PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation
2. A High-Quality and Large-Scale Dataset for English-Vietnamese Speech Translation
3. A Vietnamese-English Neural Machine Translation System
4. Recent Developments on Espnet Toolkit Boosted By Conformer
5. Librispeech: An ASR corpus based on public domain audio books
6. A study of BPE-based language modeling for open vocabulary Latin language OCR
8. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions
9. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis
10. Multilingual Denoising Pre-training for Neural Machine Translation
11. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit
12. The Stanford CoreNLP Natural Language Processing Toolkit
13. Parallel corpora for medium density languages
14. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora
15. Iterative, MT-based Sentence Alignment of Parallel Texts
References

17. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia
18. MuST-C: A multilingual corpus for end-to-end speech translation
19. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation
20. Gentle forced aligner
22. End-to-End Speech Translation with Pretrained Models and Adapters: UPC at IWSLT 2021
Standard encoder-decoder Transformer-based sequence-to-sequence models, pre-trained on large-scale corpora with a denoising objective, e.g. BART, T5, ByT5.