Recent Advances in English-Vietnamese Text and Speech Translation

Dat Quoc Nguyen

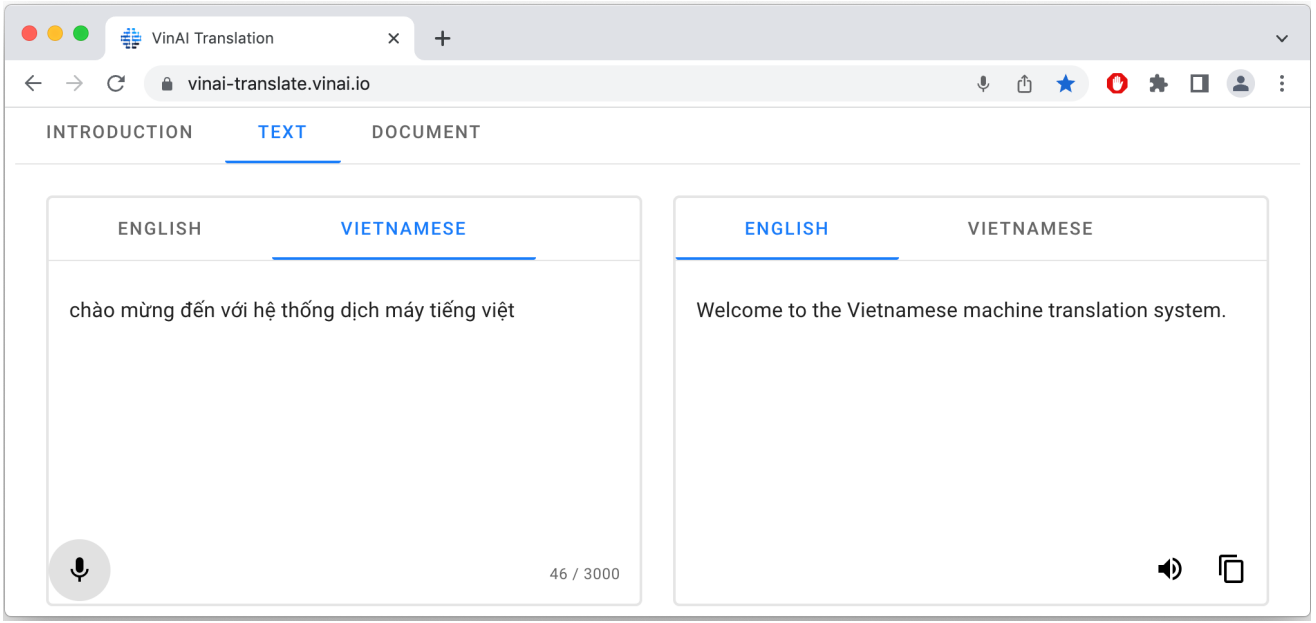
Head of NLP, VinAI

https://datquocnguyen.github.io

12/10/2022

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



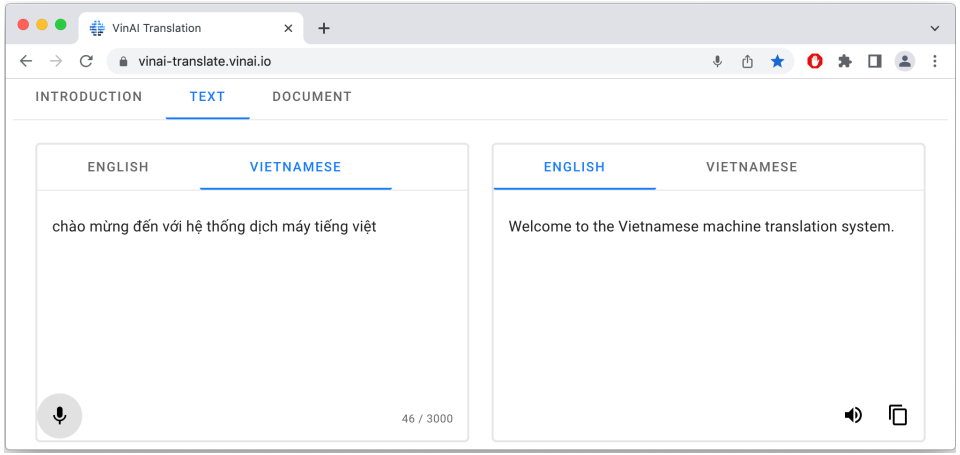
Motivation

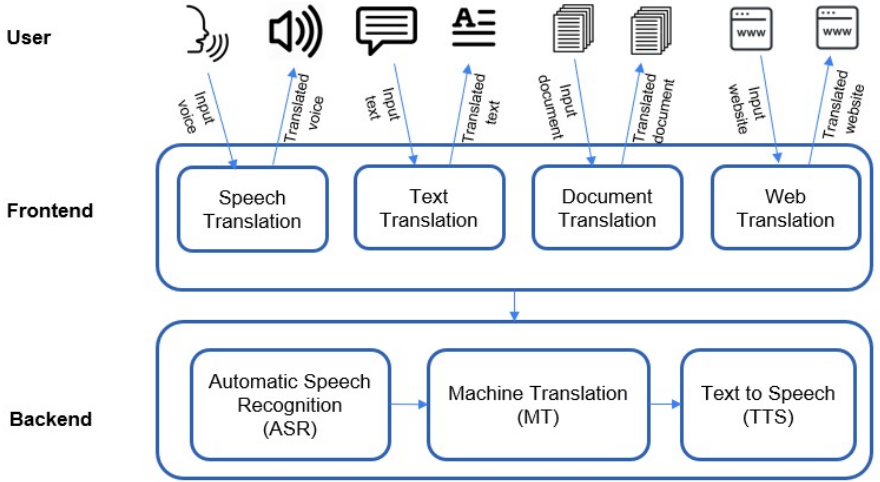
- Issues:
  - High-quality Vietnamese-English parallel corpora are either not publicly available or small-scale
  - Larger Vietnamese-English parallel corpora are noisy
- Approach:
  - Construct high-quality and large-scale parallel datasets
    - PhoMT: A High-Quality and Large-Scale Benchmark Dataset 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





Automatic Speech Recognition

- For Vietnamese
  - Train Conformer-CTC [4] using an in-house 5700-hour dataset augmented by noise injection and intensity adjustment approaches
  - 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

Domain	Total		Training			Validation			Test		
	#doc	#pair	#pair	#en/s	#vi/s	#pair	#en/s	#vi/s	#pair	#en/s	#vi/s
News	2559	41504	40990	24.4	32.0	257	22.3	30.3	257	26.8	34.5
Blogspot	1071	93956	92545	25.0	34.6	597	26.4	37.8	814	23.7	31.5
TED-Talks	3123	320802	316808	19.8	23.8	1994	20.0	24.6	2000	22.0	27.9
MediaWiki	38969	496799	490505	26.0	32.8	3024	25.3	32.3	3270	27.0	33.7
WikiHow	6616	513837	507379	18.9	22.4	3212	17.9	21.5	3246	17.5	21.5
OpenSub	3312	1548971	1529772	9.7	11.1	9635	9.5	10.7	9564	10.0	11.4
All	55650	3015869	2977999	15.7	19.0	18719	15.3	18.7	19151	16.2	19.8

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 CCAligned [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

Model	#params	Max length
vinai/vinai-translate-vi2en	448M	1024
vinai/vinai-translate-en2vi	448M	1024

- 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
- Users can also try these models at: https://huggingface.co/spaces/vinai/VinAI_Translate

PhoMT evaluation results

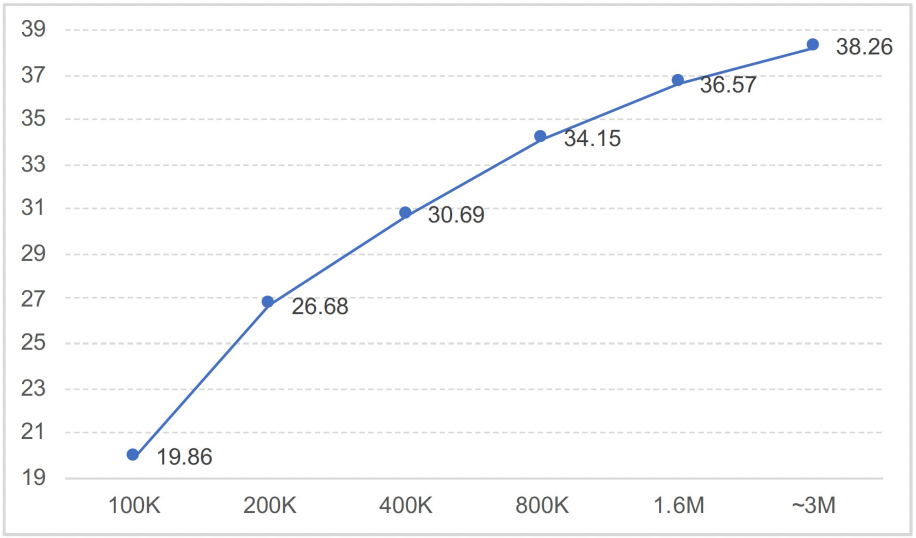
- Experimental results on the PhoMT validation and test sets while using the PhoMT training set of 2.97M pairs for training

Model	Validation set				Test set					
	En-to-Vi		Vi-to-En		En-to-Vi			Vi-to-En		
	TER↓	BLEU↑	TER↓	BLEU↑	TER↓	BLEU↑	Human↑	TER↓	BLEU↑	Human↑
Google Translate	45.86	40.10	44.69	36.89	46.52	39.86	23/100	45.86	35.76	10/100
Bing Translator	45.36	40.82	45.32	36.61	46.04	40.37	14/100	46.09	35.74	15/100
Transformer-base	42.77	43.01	43.42	38.26	43.79	42.12	13/100	44.28	37.19	13/100
Transformer-big	42.13	43.75	43.08	39.04	43.04	42.94	18/100	44.06	37.83	28/100
mBART	**41.56**	**44.32**	**41.44**	**40.88**	**42.57**	**43.46**	**32/100**	**42.54**	**39.78**	**34/100**

- 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

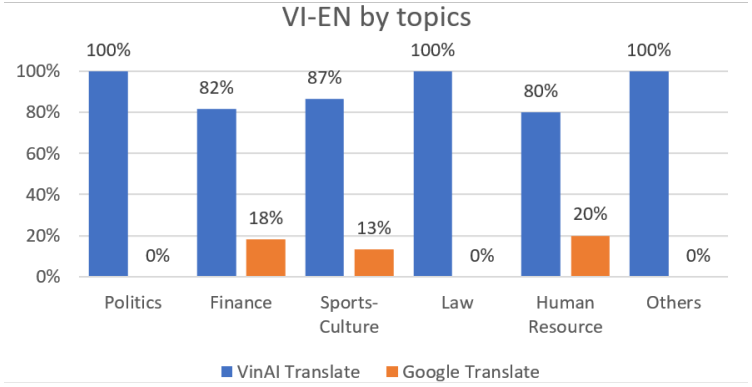
Our curation effort paid off!

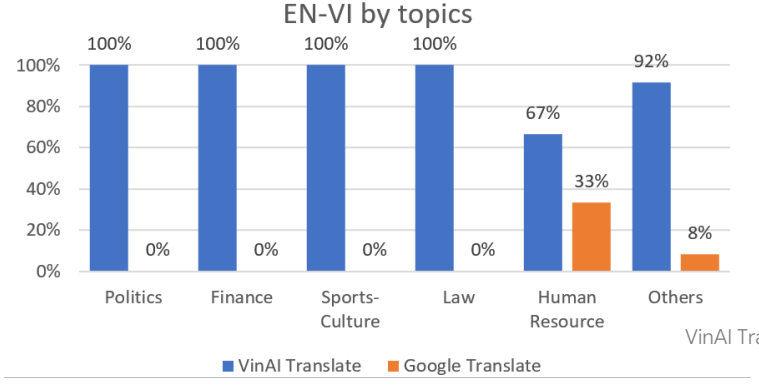
VinAI Translate evaluation results

- Automatic evaluation results

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

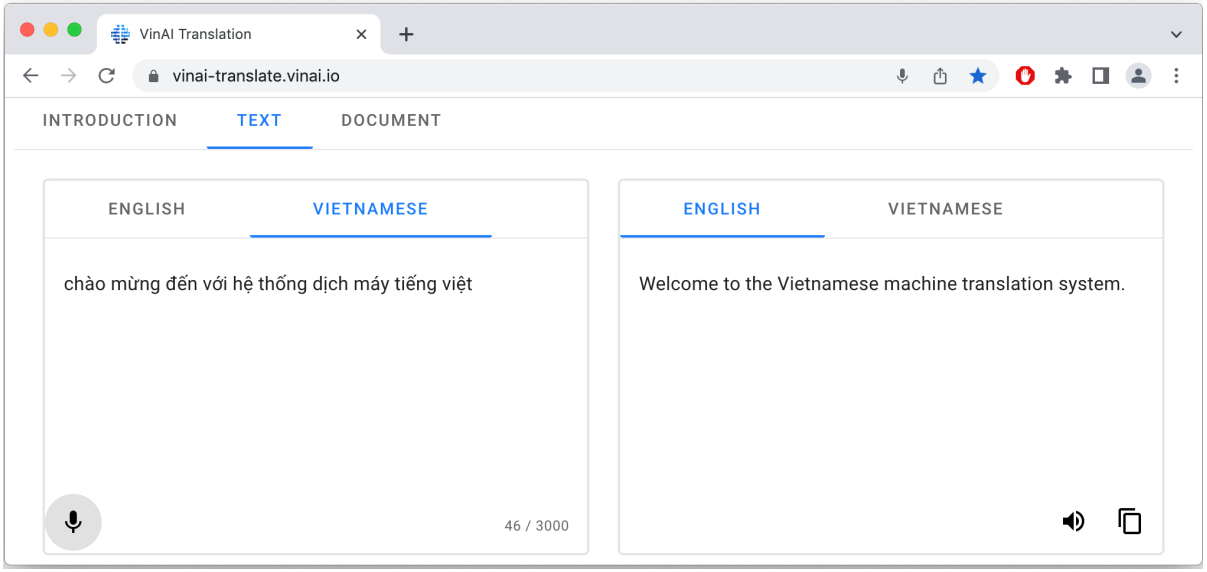
- Human evaluation results





Takeaways

- PhoMT—A high-quality and large-scale Vietnamese-English parallel dataset: https://github.com/VinAIResearch/PhoMT
- State-of-the-art translation models pre-trained for Vietnamese-to-English and English-to-Vietnamese: https://github.com/VinAIResearch/VinAI_Translate



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

Split	#triplets	#hours	#en/s	#vi/s
Training	327370	501.59	16.55	20.94
Validation	1933	3.13	17.24	22.22
Test	1976	3.77	19.23	25.65

  - #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

      ⇒ 327370 * 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

Model		BLEU↑
Casc.	(I) mBART w/ our extended dataset	33.65
	(II) mBART w/ PhoMT combination	**34.31**
E2E	S2T Transformer	29.98
	UPC's Adaptor	**33.30**

- (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

Training data	WER↓	BLEU↑
MuST-C En-Vi training set	9.09	31.66
Our sampled training subset	7.44	32.37

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: "*Nhưng trên sóng biển dài, bạn sẽ lăn dọc theo, thư giãn, năng lượng thấp*."
  - Output of the end-to-end Adaptor model trained on the subset of the PhoST training set: "*Nhưng trên một vùng biển có bước sóng dài, bạn sẽ lăn dọc, thư giãn, ít tốn 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
7. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search
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

16. CCAligned: A Massive Collection of Cross-Lingual Web-Document Pairs
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
21. Fairseq S2T: Fast Speech-to-Text Modeling with Fairseq
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

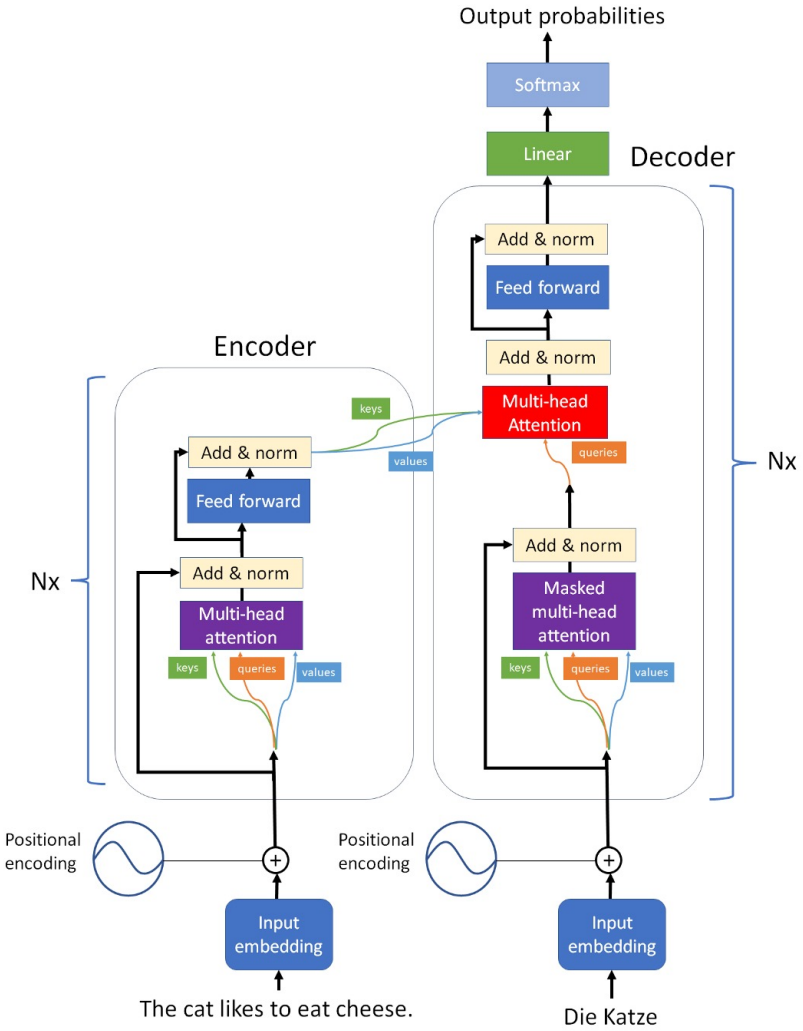


Image by Vincent Mueller: https://medium.com/towards-data-science/attention-please-85bd0abac41