Machine Translation

English to Vietnamese

Group Members Contribution

Hoàng Trung Khải 20225502 – T5 models, UI

Lưu Thiện Việt Cường 20225477 – LSTM with attention

Nguyễn Thành Minh 20225450 – LSTM without attention

Nguyễn Việt Anh 20225434 – GRU with attention

Trịnh Duy Phong 20220065 – BERT-BARTpho, MarrianMT

Table of contents

01

Introduction

Machine Translation

Machine translation is the process of using artificial intelligence to automatically translate text from one language to another without human involvement. Modern machine translation goes beyond simple word-to-word translation to communicate the full meaning of the original language text in the target language.

02

Dataset

PhoMT Dataset

PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine Translation

December 20, 2021

Motivation

Vietnam has achieved rapid economic growth in the last two decades. It is now an attractive destination for trade and investment. Due to the language barrier, foreigners usually rely on automatic machine translation (MT) systems to translate Vietnamese texts into their native language or another language they are familiar with, e.g. the global language English, so they could quickly catch up with ongoing events in Vietnam. Thus the demand for high-quality Vietnamese-English MT has rapidly increased. However, state-of-the-art MT models require high-quality and large-scale corpora for training to be able to reach near human-level translation quality. Despite being one of the most spoken languages in the world with about 100M speakers, Vietnamese is referred to as a low-resource language in MT research because publicly available parallel corpora for Vietnamese in general and in particular for Vietnamese-English MT are not large enough or have low-quality translation pairs, including those with different sentence meanings (i.e. misalignment).

Overall

3 minutes

Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, Dat Quoc Nguyen

Share Article

PhoMT Dataset

Our contributions

- We present **PhoMT**, a high-quality and large-scale Vietnamese-English parallel dataset, consisting of 3.02M sentence pairs.
- We empirically investigate strong neural MT baselines on our dataset and compare them with well-known automatic translation engines.
- We publicly release our PhoMT dataset for research or educational purposes.

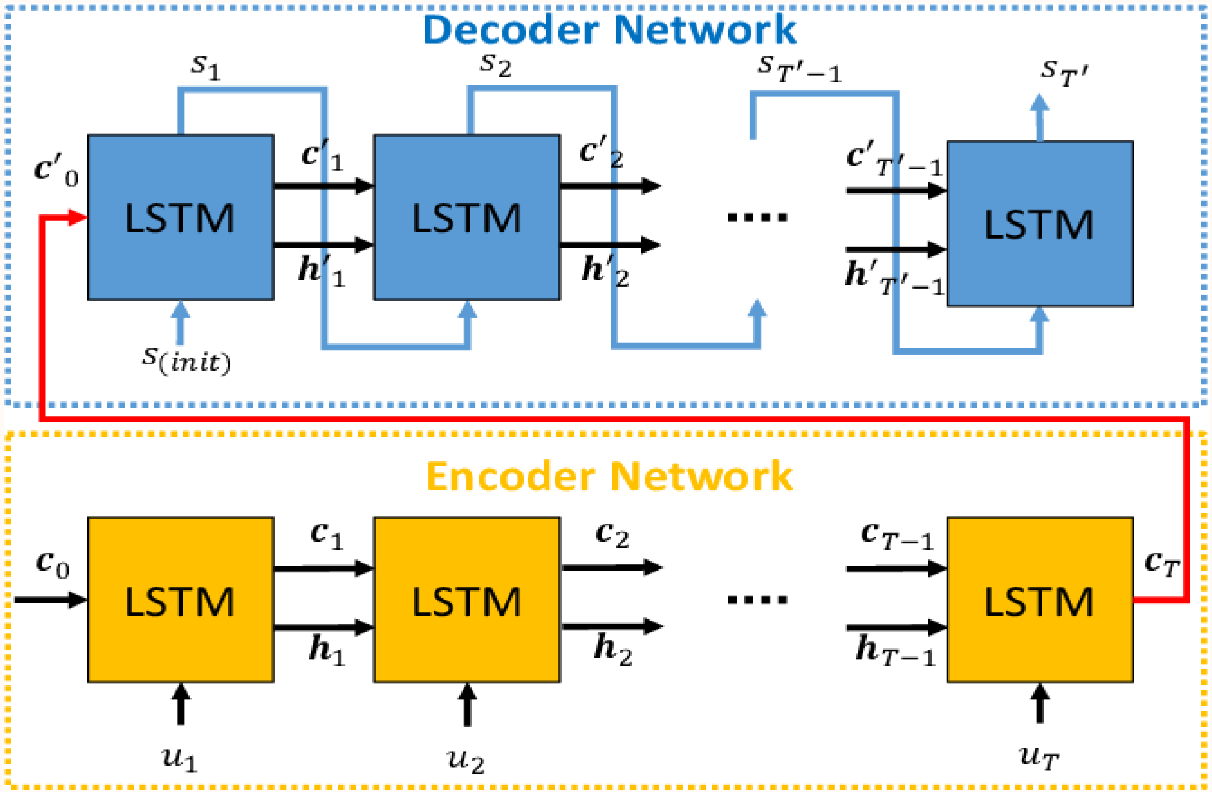
Tokenizer

1. BERT
2. BARTpho-word
3. PhoBERT base
4. T5-small
5. SentencePiece
6. Tokenizer of Keras

03

Models

Bidirectional-LSTM without Attention



Bidirectional-LSTM without Attention

Architecture:

- **Encoder:** Bi-LSTM layers-process input
- **Decoder:** Undirectional LSTM layers-generates the target sequence

Parameters:

Encoder:
– input dim = 27493
– embedding dim = 256
– units = 128
– input length = 193
– Drop out = 0.2
– Regularizer = 0.0001

Decoder:
– output dim = 15289
– embedding dim = 256
– units = 128
– input length = 232
– Drop out = 0.2
– Regularizer = 0.0001

Bidirectional-LSTM without Attention

Tokenization:

- English Tokenization: The Tokenizer from Keras is used
- Vietnamese Tokenization: The Tokenizer from Keras is used
- Sequence Conversion + Padding

Dataset:

- **Dataset Overview**: A parallel corpus of over 23,000 English-Vietnamese sentence pairs
- **Text Cleaning**: A custom clean_text function processes sentences to retain lowercase letters, Vietnamese diacritics, and specific characters (', ., ,), while handling consecutive and leading/trailing whitespaces.

Bidirectional-LSTM with attention

Architecture:

- **Encoder:** Stack of Bi-LSTM layers-process input
- **Decoder:** Stack of Bi-LSTM layers-generates the target sequence

Parameters:

Encoder:
– input dim = 27493
– embedding dim = 256
– units = 128
– input length = 193
– Drop out = 0.2

Decoder:
– output dim = 15289
– embedding dim = 256
– units = 128
– input length = 232
– Drop out = 0.2
– Number of head = 2

Bidirectional-LSTM with attention

Data

- 2% of the original PhoMT dataset for training and validating

Tokenization:

- English Tokenization: bert-base-uncased
- Vietnamese Tokenization: vinai/phobert-base
- Sequence Conversion + Padding

Bidirectional-LSTM with attention

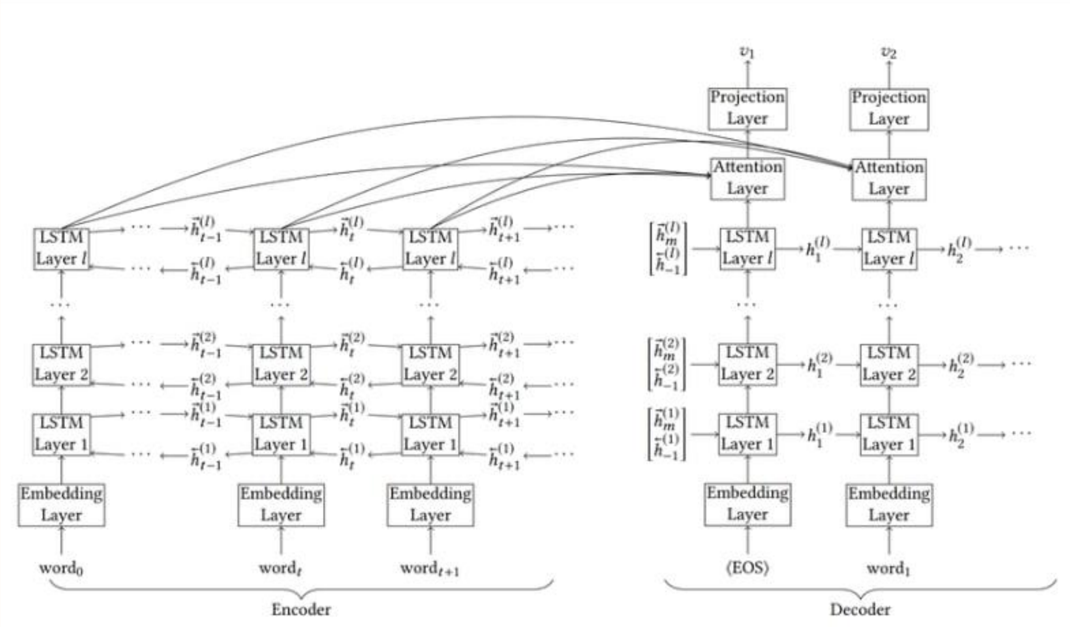


Figure 1: A Bidirectional Multilayer LSTM Encoder and UnidirectionalLSTM Decoder (Source: Yin et al.)

Bidirectional-GRU with attention

Architecture:

- **Base Tech**: Employs GRUs with an attention mechanism to enhance performance in multilingual tasks
- **Encoder**: Stack of Bi-GRU to process the input sequence
- **Attention**: Cross Attention Mechanism
- **Decoder**: Stack of GRU layers to-generate the target sequence

Bidirectional-GRU with attention

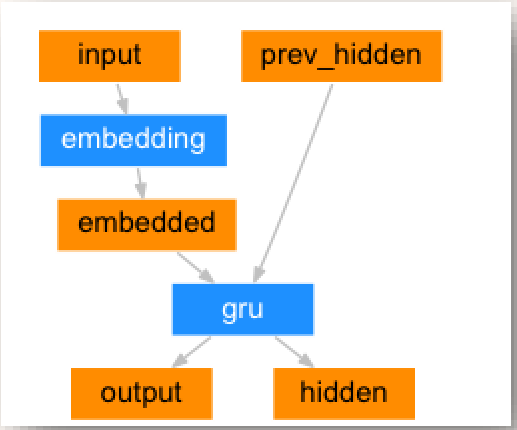


Figure 1: Encoder structure

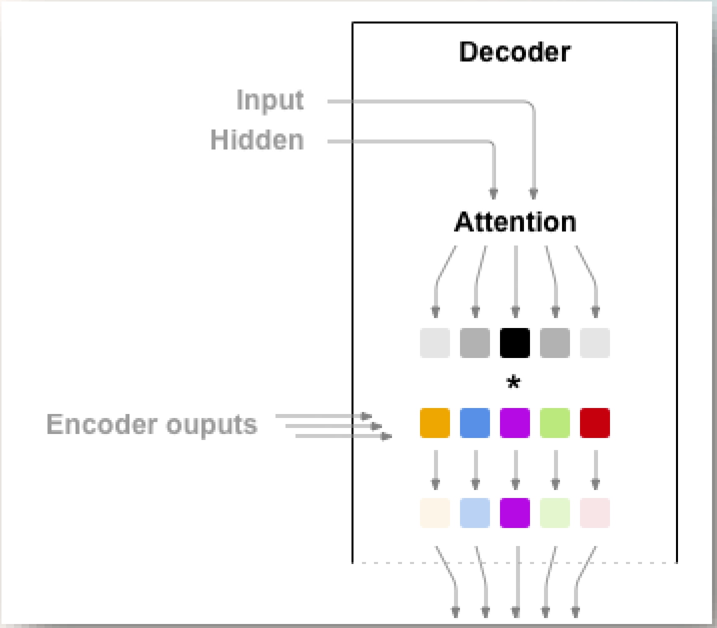


Figure 2: Decoder structure with attention

Bidirectional-GRU with attention

Dataset:

- 2% of PhoMT Dataset
- Training Data: 80% of the dataset
- Validating Data: 20% of the dataset

Tokenizer:

- tokenizer_en: BERT-base-uncased
- tokenizer_vi: PhoBERT-base

Bidirectional-GRU with attention

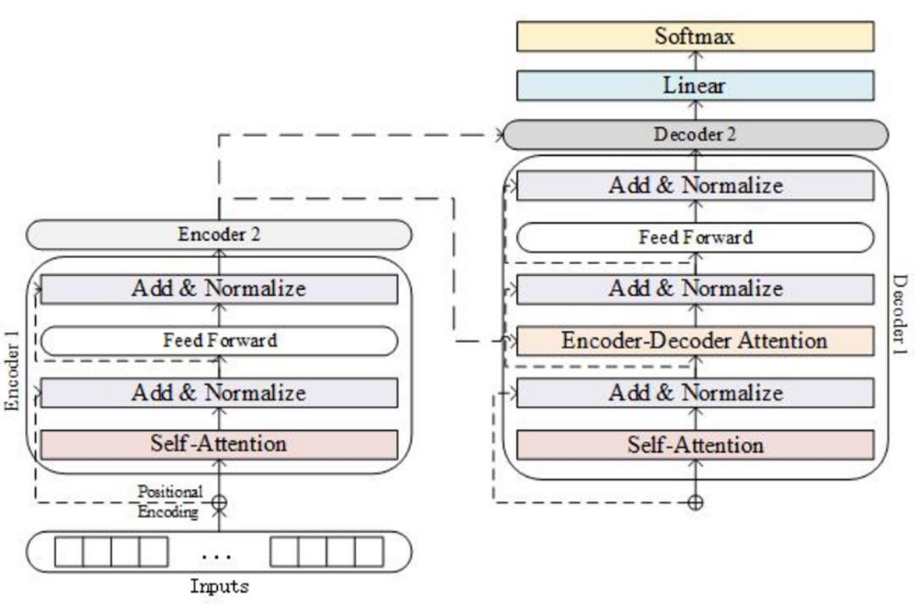
Training Configuration:

- batch_size: 32
- learning_rate: 0.001
- num_epoch: 25
- patience: 3 (early stopping)
- num_heads: 2 (for multihead attention)

T5 Model

Architecture:

- Pre-trained transformer architecture developed by GoogleResearch



T5 Model

Dataset:

- Phase 1: Trained on 20,000 sentence pairs, 4000 sentences for validation
- Phase 2: Trained on 80,000 sentence pairs, 8000 sentences for validation
- Phase 3: Trained on 200,000 sentence pairs, 8000 sentences for validation

Tokenizer:

- en_tokenizer : T5Tokenizer.pretrained("t5-small")
- vi_tokenizer : SentencePiece with the same configuration as the T5Tokenizer

T5 Model

Finetune Configuration:

- Warmup Steps: 500
- Learning Rate: 5e-4
- Batch Size: 128 for both training and 64 for validation
- Maximum Sequence Length: 64 tokens
- Epochs: total of 30 epochs
- Early Stopping: Implemented with a patience of 3 epochs to prevent overfitting
- Weight Decay: 0.1 to prevent overfitting
- Scheduler: Linear learning rate scheduler
- Evaluation Metric: Validation loss (eval_loss) was used to select the best model.
- Checkpoints: Training resumes from the last checkpoint if available.
- Number of Workers: 4 data loader workers to speed up data loading.

Bert and Bartpho-Word encoder-decoder

Architecture:
- Encoder: Bert
- Decoder: Bartpho-Word

Data:
- 297,799 data samples

Training:
- 10 training epochs

MarianMT

Model name: Helsinki-NLP/opus-mt-en-v

Data:
- 297,799 data samples

Training:
- 5 example

04

Results

Evaluation metrics- BLEU metric

- **BLEU (BiLingual Evaluation Understudy)** measures the similarity between machine-translated text and reference translations, scoring from 0 (no overlap, low quality) to 1 (perfect overlap, high quality).
- **BLEU_N**: In the BLEU metric, **BLEU_N** refers to the evaluation based on n-grams of size N, comparing how closely n-grams in the machine output match those in the reference text.
- In this evaluation step: we use range of:
  - BLEU-1: Uses **unigrams (1-grams)** to assess word-level matches.
  - BLEU-2: Considers **bigrams (2-grams)** to measure contextual and syntactic correctness.
  - BLEU-3: Evaluates **trigrams (3-grams)** for more extended phrase consistency.
  - BLEU-4: Uses **4-grams**, capturing higher-level coherence.

Evaluation metrics- Cosine Similarity

- **Cosine Similarity** is a metric that measures the similarity between two zero vectors in a multidimensional space. It calculates the cosine of the angle between the vectors, which indicates their directional similarity.
- With consine similarity metric, we can capture the segmantic meaning of the predictions of our model and references. So, this can address the issue of consider only word by word of **BLEU**
- In evaluation phases, we use the model pretrain **BartPho-word** to generate the vector of predictions and references to calculating the similarity of two sentences.

Evaluation scores

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	Cosine Similarity
LSTM without attention	0.18	0.07	0.05	0.03	0.57
GRU with attention	0.2	0.08	0.06	0.03	0.59
LSTM with attention	0.31	0.17	0.13	0.07	0.66
T5	0.33	0.23	0.19	0.11	0.64
BERT – BARTpho	0.56(0.03)	0.44(0.01)	0.39(0.01)	0.30(0.13)	0.82(0.29)
MarrianMT	0.61(0.39)	0.48(0.29)	0.44(0.24)	0.33(0.16)	0.84(0.73)

Thanks!