Automatic Post-Editing for Vietnamese

Thanh Vu, thanh.v.vu@oracle.com

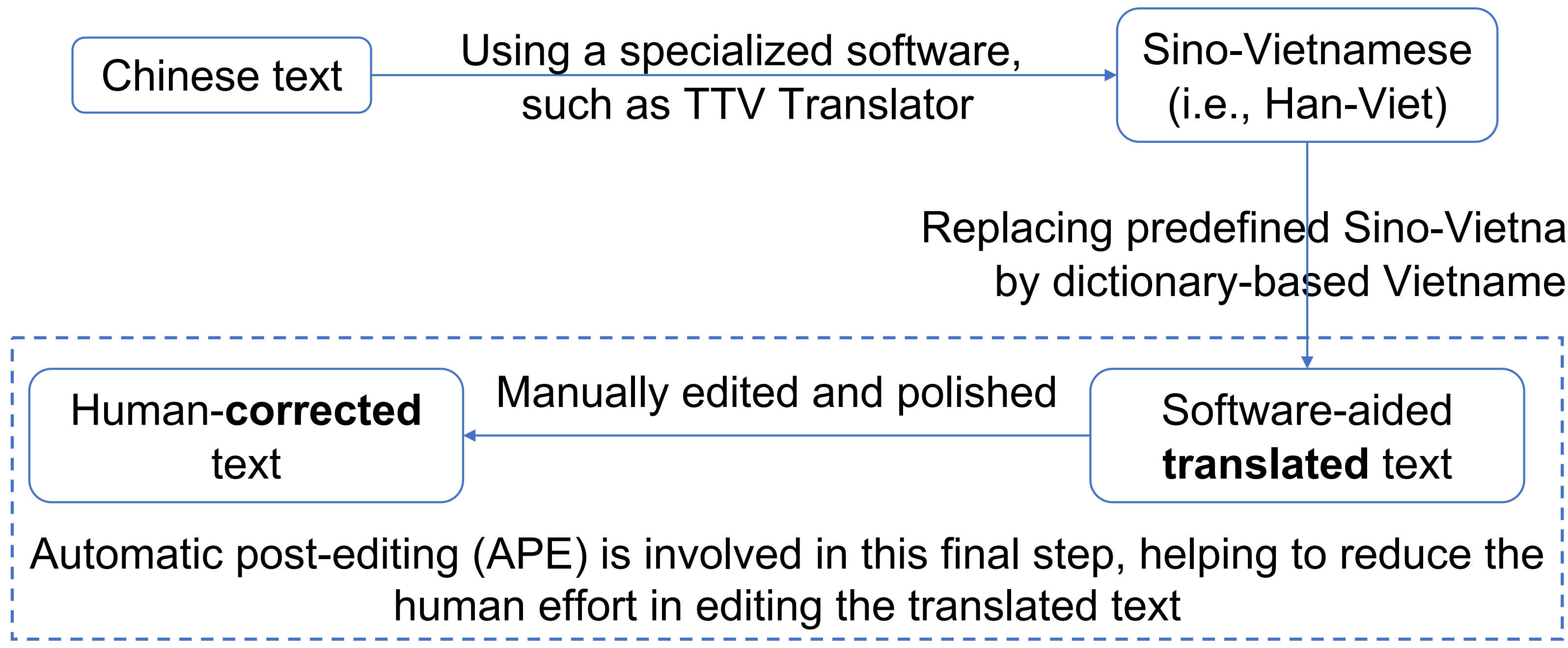
Dai Quoc Nguyen, dai.nguyen@oracle.com

ALTA 2021

Introduction

Translating Chinese novels to Vietnamese is an important task. In the last ten years, there are about 30K Chinese novels describing fiction stories, that are available in Vietnamese with ~80K active readers and ~600K novel chapter views daily from the three most popular Vietnamese websites for reading novels.

But, translating the Chinese novels to Vietnamese is still challenging. The reason is that in fact, readers prefer reading the novels translated using the traditional language style rather than the modern language style used in news articles. Note that current general-purpose MT systems (e.g., Google Translate), trained on modern language style-focused bilingual corpora, cannot satisfy the reader preference.



Chinese text

Using a specialized software, such as TTV Translator

Sino-Vietnamese (i.e., Han-Viet)

Replacing predefined Sino-Vietnamese phrases by dictionary-based Vietnamese phrases

Software-aided **translated** text

Manually edited and polished

Human-**corrected** text

Automatic post-editing (APE) is involved in this final step, helping to reduce the human effort in editing the translated text

- We formulate the APE problem for Vietnamese as a monolingual translation task. We are the first to tackle the APE task for Vietnamese to automatically improve the quality of the Vietnamese translated text of Chinese novels.

- We create a large-scale dataset of 5M translated and corrected sentence-level pairs extracted from 99.5K translated and corrected chapter-level pairs from 183 novels.

- We empirically evaluate neural MT models using our dataset, including a fully convolutional model (Gehring et al., 2017), "Transformer-base" and "Transformer-large" (Vaswani et al., 2017). We compare these models under automatic- and human-based evaluation settings as well as in-domain and out-of-domain schemes.

- We publicly release our dataset and model checkpoints (for research-only purpose) at: https://github.com/tienthanhdhcn/VnAPE

Our dataset

Dataset construction

- Of 30K Chinese novels available in Vietnamese, there are currently only 283 novels available in both Vietnamese translated and corrected texts. We crawl all of those 283 novels.
- There is a ground-truth chapter-level alignment between translated and corrected chapter-level pairs from each of the 283 novels.
- We randomly sample from each novel 5 pairs of translated and corrected chapters and employ three annotators to manually evaluate the sampled chapters' editing quality on a 5-point scale.
- We select the top 183 novels having the highest average points over their sampled chapters to be included in our dataset.
- We use all translated and corrected chapter-level pairs from the top 183 novels, i.e., a total of 99.5K chapter-level pairs. We then use RDRSegmenter (Nguyen et al., 2018) from VnCoreNLP (Vu et al.,2018) to segment each chapter text into individual sentences.
- In each chapter, to align the translated and corrected sentences, we compute an alignment score. In the end, our dataset consists of 5M (i.e., 5,028,749) translated and corrected sentence-level pairs in Vietnamese.

Dataset splitting

- For the in-domain scheme, the dataset is split based on the novel chapters, in which the first 92.5% chapters of each novel are used for training, the next 2.5%are for validation, and the last 5% are for testing.
- For the out-of-domain scheme, we split our dataset into training, development and test sets such that no novel overlaps between them. We select novels for training, validation and test sets so that the out-of-domain data distribution is similar to thein-domain data distribution.

Experimental results

Automatic evaluation

- Regarding the in-domain scheme, the neural MT models produce substantially higher GLEU and BLEU scores and a lower TER score than the translated text. This indicates that APE helps improve the quality of the translated text.

Item	Training set		Validation set		Test set	
	Translated	Corrected	Translated	Corrected	Translated	Corrected
#chapters(#novels)	92.2K (183)		2.5K (183)		4.8K (183)	
#sentences	4.65M		126.7K		248.0K	
#tokens	152.1M	143.7M	4.1M	3.9M	8.1M	7.6M
#tokens/sentence	32.7	30.9	32.7	31.0	32.6	30.8

Table 1: In-domain statistics of our dataset.

Item	Training set		Validation set		Test set	
	Translated	Corrected	Translated	Corrected	Translated	Corrected
#chapters(#novels)	91.5K (128)		2.8K (28)		5.1K (27)	
#sentences	4.66M		120.1K		245.6K	
#tokens	151.3M	143.0M	4.1M	3.8M	8.9M	8.4M
#tokens/sentence	32.5	30.7	33.7	31.6	36.3	34.2

Table 2: Out-of-domain statistics of our dataset.

Model	In-domain			Out-of-domain		
	TER↓	GLEU↑	BLEU↑	TER↓	GLEU↑	BLEU↑
translated	46.027	39.816	35.834	50.678	36.174	31.591
fconv	36.539	49.188	47.933	43.106	42.654	40.502
Transformer-base	35.882	49.803	48.588	42.970	42.726	40.588
Transformer-large	**35.161**	**50.763**	**49.686**	**42.892**	**42.818**	**40.704**

Table 3: Experimental results on the test sets. "translated" denotes the result computed in using the raw translated sentence without post-editing correction.

- Regarding the out-of-domain scheme, all three neural MT models help improve the quality of the translated text with the absolute improvements of at least 7.5, 6.5, 9.0 points for TER, GLEU,BLEU, respectively.

Human evaluation

- We conduct a human evaluation to manually evaluate the output quality of the three trained models. In particular, we collect a new set of 1K translated sentences which are randomly selected from 10 novels that are not in our dataset. To perform APE, we then apply each of the three models to produce a ``corrected'' candidate output for each ``translated'' sentence, resulting in three corrected candidates.
- We ask three annotators to independently vote the most suitable sentence among the translated sentence and its three corresponding corrected candidates (here, we do not show which sentence is the translated one or corrected by which model to the annotators).
- The results for the human evaluation are consistent with the results produced by the three models on the test sets, confirming the effectiveness of ``Transformer-large'' for APE in Vietnamese.