Automatic Post-Editing for Vietnamese

Thanh Vu[1] and **Dai Quoc Nguyen**[2*]
[1]Oracle Digital Assistant, Oracle, Australia
[2]Oracle Labs, Oracle, Australia
[1]thanh.v.vu@oracle.com; [2]dai.nguyen@oracle.com

Abstract

Automatic post-editing (APE) is an important remedy for reducing errors of raw translated texts that are produced by machine translation (MT) systems or software-aided translation. In this paper, we present a systematic approach to tackle the APE task for Vietnamese. Specifically, we construct the first large-scale dataset of 5M Vietnamese translated and corrected sentence pairs. We then apply strong neural MT models to handle the APE task, using our constructed dataset. Experimental results from both automatic and human evaluations show the effectiveness of the neural MT models in handling the Vietnamese APE task.

1 Introduction

Recent research has placed significant advancements for automatic machine translation (Wu et al., 2016; Vaswani et al., 2017; Barrault et al., 2019). The high-quality MT output has been widely adopted by professional translators into their translation workflow to save time and reduce translation errors (Zaretskaya et al., 2016).

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 $\sim$ 80K active readers and $\sim$ 600K novel chapter views daily from the three most popular Vietnamese websites for reading novels.[1][2][3] 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 (e.g. using "**tiểu nữ nhi**"$_{\textbf{little girl}}$ instead of "**cô bé**"$_{\textbf{little girl}}$). Note that current general-purpose MT systems (e.g., Google Translate), trained on modern language style-focused bilingual corpora, cannot satisfy the reader preference.

The well-known workflow/guideline used for translating the Chinese novels to Vietnamese consists of three steps:[4]

- In the first step, the Chinese text is converted into Sino-Vietnamese (i.e. Han-Viet)[5] text using a specialized software, such as TTV Translator.[6]
- In the second step, the Sino-Vietnamese text is further smoothed by replacing predefined Sino-Vietnamese phrases by dictionary-based Vietnamese phrases. The core content of the Vietnamese text generated as the output of the second step—namely software-aided **translated** text—can be generally understood by frequent readers who are familiar with reading the translated text. Note that the translated text does not fully follow the Vietnamese grammar and vocabulary, thus making it hard for new readers (and even fairly often for the frequent readers) to understand details of the text content.
- In the final step, the translated text is manually edited and polished following Vietnamese vocabulary and grammar. Here, we refer to the text generated as the output of the final step as the human-**corrected** text that can be accessed easily by readers with different reading levels.

Note that the final editing step is very time-consuming due to the large amount of human-manual work. Thus automatic post-editing (APE)

[*] Most of the work was done before two authors joined Oracle.

[1] https://truyencv.com
[2] https://truyenyy.com
[3] https://truyen.tangthuvien.vn

[4] http://www.tangthuvien.vn/forum/showthread.php?t=142168&page=2
[5] https://en.wikipedia.org/wiki/Sino-Vietnamese_vocabulary
[6] https://play.google.com/store/apps/details?id=vn.tangthuvien.ttvtranslate&hl=en_AU.

might be involved in this final step, helping to reduce the human effort in editing the translated text (Tatsumi, 2010). To the best of our knowledge, there is no previous study on APE for Vietnamese.

In this paper, we formulate the APE problem for Vietnamese as a monolingual translation task. We first construct a large-scale dataset consisting of translated and corrected sentence pairs. We then use our dataset to train a state-of-the-art neural MT model to automatically post-edit the translated sentences, and compare these models under various settings. Our contributions are summarized as:

- 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.

2 Our dataset

This section presents our large-scale dataset for the Vietnamese APE task.

Dataset construction

In almost all cases, the original Chinese novels are not publicly available to the readers of the Vietnamese websites for reading novels, thus *we cannot access those Chinese novels' texts*. 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 $\alpha = \frac{2 \times |I|}{|T| + |C|}$, where $|T|$ and $|C|$ denote the numbers of tokens in the translated and corrected sentences, respectively, while $|I|$ denotes the size of the intersection between them. Our sentence alignment process has two phases:

- In the first phase, we align every translated and corrected sentence pair with a score $\alpha >= 0.75$, i.e. alignment mode 1–1.
- In the second phase, for the remaining sentences, using a threshold $\alpha >= 0.5$, we only consider two alignment modes 1–2 and 2–1 for one translated sentence aligning two adjacent corrected sentences and two adjacent translated sentences aligning one corrected sentence, respectively.[7]

The alignment modes 1–1, 1–2 and 2–1 account for about 98% of the validation set.[8] In the end, our dataset consists of 5M (i.e. 5,028,749) translated and corrected sentence-level pairs in Vietnamese.

Dataset splitting

Our dataset of 5M Vietnamese translated and corrected sentence pairs is split into training, validation and test sets. We propose two splitting schemes which are *in-domain* and *out-of-domain*. 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 the in-domain data distribution. Basic in-domain and out-of-domain data statistics are detailed in tables 1 and 2, respectively.

3 Experimental setup

This section presents neural MT models as well as their training details that we employ for evaluation.

[7] We concatenate two adjacent sentences into a single one.

[8] We do not include the remaining 2% unaligned sentences into our dataset.

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.

Neural MT models

We formulate the final step of editing and polishing (i.e. post-editing) the translated sentence as a (monolingual) translation task. In particular, the translated and corrected sentences are viewed as the ones in the source and target languages, respectively. We employ strong neural MT models to handle the task. The first model is the well-known Transformer, in which we use its two variants of **"Transformer-base"** and **"Transformer-large"** (Vaswani et al., 2017). The second model is a fully convolutional model, named **"fconv"**, consisting of a convolutional encoder and a convolutional decoder (Gehring et al., 2017).

Training details

For each dataset splitting scheme, we train the models on the training set using implementations from the `fairseq` library (Ott et al., 2019). For each model, we employ the same model configuration as detailed in the corresponding paper (Vaswani et al., 2017; Gehring et al., 2017). We train each model with 100 epochs with the beam size of 5. We use the same shared embedding layer for both the encoder and decoder components of a neural MT model as both the translated and corrected sentences are in Vietnamese. We apply early stopping when no improvement is observed after 5 continuous epochs on the validation set. The model obtaining the highest BLEU score (Papineni et al., 2002) on the validation set is then used to produce the final scores on the test set.

We use standard MT evaluation metrics including TER—Translation Edit Rate (Snover et al., 2006), GLEU—Google-BLEU (Wu et al., 2016) and BLEU, in which lower TER, higher GLEU, higher BLEU indicate better performances.

4 Main results

Automatic evaluation

Table 3 shows in-domain and out-of-domain results for each model as well as for the translated text. In particular, with 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. Among the MT models, "Transformer-large" achieves the best performance with the BLEU score of 49.686 which is 1.098 and 1.753 higher than "Transformer-base" and "fconv", respectively.

Regarding the out-of-domain scheme, Table 3 also shows a similar trend. In particular, 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. We also note that although "Transformer-large" consistently achieves the best TER, GLEU and BLEU scores, the out-of-domain score differences between the neural MT models are not as substantial as in the in-domain scheme.

Human evaluation

To better understand the performances of neural MT models, 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

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.

models to produce a "corrected" candidate output for each "translated" sentence, resulting in three corrected candidates.[9]

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), thus resulting in 3,000 votes in total. The best model is "Transformer-large" obtaining 1,405 votes (46.8%), compared to 815 votes (27.2%) for "Transformer-base", 780 votes (26.0%) for "fconv" and 0 vote for the translated sentences. We measure the inter-annotator agreements between the three annotators using Fleiss' kappa coefficient (Fleiss, 1971). The Fleiss' kappa coefficient is obtained at 0.350 which can be interpreted as *fair* according to Landis and Koch (1977). 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.

5 Related work

Our work is the first one to automatically handle the task of correcting the Vietnamese translated text of Chinese novels. However, APE is not new and has proved to be an effective approach to handle the inaccuracies of raw MT output (Simard et al., 2007; Lagarda et al., 2009; Pal et al., 2016; Nguyen et al., 2017; Correia and Martins, 2019).

APE approaches cover two main research directions including statistical MT-based models (Simard et al., 2007; Lagarda et al., 2009) and neural MT-based models (Pal et al., 2016; Correia and Martins, 2019). In particular, Simard et al. (2007) propose a statistical phrase-based MT system to post-edit the output of a rule-based MT system by handling the typical errors made by the rule-based one. Likewise, Lagarda et al. (2009) utilize statistical information from a pre-trained statistical MT model to post-edit the output of another statistical MT model. Pal et al. (2016) propose to use Bidirectional LSTM encoder-decoder for APE and found that it performs better than statistical phrase-based APE. Correia and Martins (2019) present an effective APE approach where they fine-tune pre-trained BERT models (Devlin et al., 2019) on both the BERT-based encoder and decoder.

6 Conclusion

We have presented the first work of APE for Vietnamese to automatically correct the Vietnamese translated text of Chinese novels. We construct the first large-scale dataset of 5M translated and corrected sentence-level pairs, extracted from 99.5K translated and corrected chapter-level pairs from 183 novels, for the Vietnamese APE task. We then compare three MT models using our dataset under in-domain and out-of-domain data splitting schemes. Experimental results from both the automatic and human evaluations show that the neural MT models help improve the quality of the translated text. Specifically, "Transformer-large" achieves the best performances w.r.t. the TER, GLEU, BLEU scores and human votes, helping to reduce the human effort in editing the translated novels, and serving as a strong model for future research and applications. We also publicly release our dataset and model checkpoints (*for research-only purpose*) at: https://github.com/tienthanhdhcn/VnAPE.

Acknowledgements

We thank the three anonymous reviewers for their valuable comments and suggestions which help improve the quality of the paper. We would also like to thank Dat Quoc Nguyen and his team for their help and support.

[9]Note that we select the 1K translated sentences to ensure that the three corrected candidates are different.

References

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Gonçalo M Correia and André FT Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. *arXiv preprint arXiv:1906.06253*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Antonio-L Lagarda, Vicent Alabau, Francisco Casacuberta, Roberto Silva, and Enrique Díaz-de Liaño. 2009. Statistical post-editing of a rule-based machine translation system. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 217–220.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Dai Quoc Nguyen, Dat Quoc Nguyen, Cuong Xuan Chu, Stefan Thater, and Manfred Pinkal. 2017. Sequence to sequence learning for event prediction. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 37–42.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of LREC*, pages 2582–2587.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Midori Tatsumi. 2010. *Post-Editing Machine Translated Text in a Commercial Setting: Observation and Statistical Analysis*. Dublin City University. Faculty of Humanities and Social Science.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In *Proceedings of NAACL: Demonstrations*, pages 56–60.

Yonghui Wu, Mike Schuster, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint*, arXiv:1609.08144.

Anna Zaretskaya, Mihaela Vela, Gloria Corpas Pastor, and Miriam Seghiri. 2016. Comparing post-editing difficulty of different machine translation errors in Spanish and German translations from English. *International Journal of Language and Linguistics*, 3(3):91–100.