Goals, Challenges and Findings of the VLSP 2020
English-Vietnamese News Translation Shared Task

Thanh-Le Ha\textsuperscript{1,2}, Van-Khanh Tran\textsuperscript{2}, Kim-Anh Nguyen\textsuperscript{2}

\textsuperscript{1}Interactive Systems Lab, Karlsruhe Institute of Technology, Germany
thanh-le.ha@kit.edu

\textsuperscript{2}Speech and Language Processing Department, Vingroup Big Data Institute, Vietnam
\{v.leht6, v.khanhtv13, v.anhnk9\}@vinbigdata.org

Abstract

This paper reports the VLSP 2020 English-Vietnamese News Translation shared task, which is one of the six shared tasks organized at the seventh annual workshop on Vietnamese Language and Speech Processing (VLSP 2020). In this task, we provided parallel and monolingual data for training machine translation systems translating English texts into Vietnamese, with the focus of news domain. There were 6 teams participating into the tasks, with 13 submissions in total. We performed both automatic and human evaluations on the submissions and presented the results and our findings at the conference. We hope this would boost the research of Vietnamese machine translation community and start maintaining annual machine translation tasks at VLSP conferences.

1 Introduction

VLSP stands for Vietnamese Language and Speech Processing Consortium. It is an initiative to establish a community working on speech and text processing for the Vietnamese language. The VLSP 2020 is the seventh annual international workshop and evaluation campaign.

Machine Translation (MT) is one of the six shared tasks for the VLSP evaluation campaign this year and it is the first time that MT is organized as a VLSP shared task after being a trial task in 2013. As an important research problem of Language and Speech Processing (LSP), MT often attracts interests from the research community. However, research in Vietnamese language-related MT often conducted by several R&D departments from big companies and research labs in large universities. In 2015, the prestigious MT campaign IWSLT (Cettolo et al., 2015), whose conference was organized in Da Nang, Vietnam, featured English-Vietnamese MT as one of the MT task of that year’s campaign and it has been the first and only MT evaluation featuring Vietnamese language to date. We set the following goals when organizing this VLSP 2020 MT evaluation campaign:

• Reviving a traditional task in any LSP community and making it to be a recurrent event. Encouraging research for Vietnamese-related MT and engaging researcher into solving interesting problems of MT

• Motivating the contribution of free data and basic LSP tools supporting Vietnamese-related MT research.

• Extending practical applications of MT into smart tools and workflows, e.g. developing multilingual education channels, fighting again fake news in any languages and overcoming language barrier in business, tourism, entertainment and international communication.

Concretely, we have the following contributions while organizing VLSP 2020 English-Vietnamese News Translation task:

• Crawl, collect, compile and release free parallel and monolingual datasets for training and testing English-Vietnamese MT systems\textsuperscript{1}.

• Establishing a standard benchmark for research on English-Vietnamese Translation.

• Conduct automatic and human evaluations of the participating MT systems.

This paper is organized as follows. We describe the dataset for training and testing MT systems in Section 2. Section 3 lists the participating teams and summarizes the approaches they employed in

\textsuperscript{1}The datasets are published at https://github.com/thanhleha-kit/EnViCorpora
their systems. Section 4 presents how we evaluated the translation outputs. We then show the evaluation results in Section 5. Finally, we conclude the the paper by giving our findings and drawing our future plans for the task.

2 Dataset

Although English-Vietnamese is the most popular language pair in the Vietnamese MT community, it is currently considered as “low-resource” language pair, where there are only a few public English-Vietnamese parallel corpora with adequate quality for training MT systems. They are Wikipedia articles extracted for the Asian Language Treebank project (Riza et al., 2016), mixed-domain EVB parallel corpus collected by Ngo et al. (2013), a multilingual corpus of short and basic sentences from Tatoeba project\(^2\) (Tiedemann, 2012) and a COVID-19 multilingual corpus created by ELRC\(^3\) and compiled by Tiedemann (2012). In total, those corpora contain around 75,000 English-Vietnamese sentence pairs. Besides those high quality datasets, OPUS\(^4\)(Tiedemann, 2012), a website collecting translated texts from the web, compiles and publishes clean versions of movie subtitle datasets extracted from OpenSubtitles\(^5\) (Lison and Tiedemann, 2016), as well as religious news and bible translations. Although they are large corpora with the number of sentence pairs varying from hundreds thousands to more than three millions, they are unusable without any filtering method since the domain are very narrow (religious) and the quality is not good (movie subtitles).

\(^2\)https://tatoeba.org/eng/
\(^3\)https://elrc-share.eu/
\(^4\)https://opus.nlpl.eu/
\(^5\)https://www.opensubtitles.org/en

| Dataset Name | Domain | Size (# Sentence Pairs) |
|--------------|--------|-------------------------|
| News         | News (in-domain) | 20K                     |
| Basic        | Basic and short conversation | 8.8K                   |
| EVBCorpus    | Mixed domains   | 45K                     |
| TED-like     | Educational & Tech Talks | 546K                   |
| Wiki-ALT     | Wikipedia articles | 20K                     |
| OpenSubtitle | Movie Subtitles  | 3.5M                    |
| Corpus.2M.shuf | Monolingual Corpus of Vietnamese News | 2M                      |

Table 1: *Training Datasets for VLSP 2020 MT task*

2.1 Training Data

Parallel Data. We decide to create more parallel data for English Vietnamese. Our crawling sources are high-quality bilingual or multilingual websites of news and one-speaker educational talks of various topics, mostly technology, entertainment and design (hereby referred as TED-like talks). Because those websites are required to convey the original content in English to other languages (including Vietnamese) and often gone through several review stages before publishing, the quality is assured.

First, we extracted some basic conversations from English teaching websites and coupled them to the Tatoeba dataset. For the news domain, we crawled the data from and then applied some simple filtering methods to remove short sentences. Finally, we combined the crawled data with the COVID-19 ELRC data to produce a 20,000-sentence-pair parallel corpus.

For the TED-like domain, we downloaded TED talks monolingual data of English and Vietnamese from WIT3\(^6\) (Cettolo et al., 2012), then aligned them based on the sentence ids. Furthermore, we extracted a parallel corpus from the subtitles of the TED-like videos uploaded on Amara\(^7\) - a platform to assist its users to produce captions and subtitles of the videos they uploaded. As the result, more than five hundreds thousands sentence pairs were crawled.

Since the quality of the large OpenSubtitle dataset varies in movies, we decided to include it into the training data and let the participants choose how to use it. In the end, we released the following training data in which news is the in-domain data:

Monolingual Data. For this evaluation, we provided target monolingual data which is 2 million Vietnamese sentences, crawled from Vietnamese

\(^6\)https://wit3.fbk.eu/
\(^7\)https://amara.org/en/
newspapers from various topics. The text has adequate quality to train language models or to conduct back translation. Similar to the parallel data, we let the participants decide how to preprocess the data.

2.2 Validation and Test Data

Validation Data. While crawling the news data for training, we also reserved a small part to be validation data. We released a development dataset and a public test dataset at the same time with the training data. The development set contains 1007 English-Vietnamese sentence pairs and the public test set contains 1220 English-Vietnamese sentence pairs. The participants could use one of the validation sets to turn their models’ hyperparameters and the other sets for choosing the primary system to be submitted.

Official Test Data. We informed the participants in advance that the in-domain data is News, but we did not reveal the theme is Covid-19 News until the report of the evaluation campaign. In order to avoid cheating and accidentally inclusion of the test data into training or validation data, we manually selected up-to-date English news about Covid-19 from international online newspapers and then asked professional translators to translate them into Vietnamese. The translators need to conform some strict guidelines while translating the official test set, in order to keep it high quality. As the result, the official test set contains 789 sentence pairs. We mixed them with other crawled 2000 sentence pairs and distributed the English part to the participants, asking them to produce the Vietnamese translation using their models.

3 Participants and their Approaches

The organizers received submissions from 6 different teams with the total number of 13 submissions. Table 2 lists the teams. Among them, there are only 3 teams sending their paper describing their approaches and models.

| Team               | Affiliation | Submitted |
|--------------------|-------------|-----------|
| Bluesky            | Unknown     | 2         |
| EngineMT (Ngo et al., 2020) | UET-ICTU    | 6         |
| Lab-914 (Le and Nguyen, 2020) | HUST       | 2         |
| NLP-HUST           | HUST        | 1         |
| THORLab            | D-Soft      | 1         |
| RD-VAIS (Pham et al., 2020) | TNU-HUST-VAIS | 1         |

Table 2: The teams participated to VLSP 2020 MT task

3.1 Architecture

All of the three teams submitted neural machine translation systems. And all of them implemented their systems using the state-of-the-art Transformer architecture (Vaswani et al., 2017). The configurations are different, however. In EngineMT and RD-VAIS systems, the number of layers is 4 and the model size is 512 while in Lab-914 the number of layers is 6 and the model size is 1024.

3.2 Preprocessing

In the preprocessing phase, the teams utilized common techniques on the parallel data. They all removed long sentences, tokenized the words simply by white-spaces and applied some casing treatments. In addition, Lab-914 performed those techniques plus further filtering methods to remove noisy sentences from the Vietnamese monolingual corpus. For casing treatments, Lab-914 simply lower-cased the data, RD-VAIS marked capitalized and upper-cased words by some special tokens before lowercasing and EngineMT applied smart casing using Moses toolkit (Koehn et al., 2007). All the teams performed subword tokenization using Byte-Pair Encoding algorithm (Sennrich et al., 2016b) implemented in subword-nmt framework with the number of merging operations set at 35,000.

3.3 Back Translation

All the teams employed Back Translation (Sennrich et al., 2016a) as the sole technique to exploit monolingual data. However, each team had different strategies on how to use the monolingual data. Lab-914 used all the monolingual data provided while EngineMT used much smaller monolingual corpus after filtering out most of them using their proposed data selection techniques. RD-VAIS, on the other hand, built two systems different on the
size of the monolingual corpus. One contains 1 million sentences and the other contains all 2 million sentences. At the end, they chose the system trained with 1 million sentences back translation based on the performance on the public test set.

3.4 Domain Adaptation

We organized the task in a way that we would expect to see some domain adaptation techniques. Lab-914 did not employ any specific domain adaptation when they used all the provided data in their systems and treated the in-domain news data the same as other data. RD-VAIS, besides the monolingual data which is news, they used only the parallel in-domain data. This might affect badly on their systems since the in-domain data is small and most of their training data come from back translated data. EngineMT is the team who employed several domain adaptation approaches. First they select subsets of data, both monolingual and parallel, which are relevant to the in-domain data with their TF-IDF-based data selection technique. Then they fine-tuned their models on the in-domain data and ensembled all the models they had.

4 Evaluation

VLSP 2020 is the first Machine Translation Evaluation Campaign for Vietnamese that has both automatic and human evaluation. Furthermore, the human evaluation result is used to rank the teams in the campaign.

4.1 Automatic Evaluation

For this campaign, we employed two metrics to evaluate the submissions: BLEU and TER. Since BLEU is the most popular automatic evaluation metric in Machine Translation, it is the main metric to rank submissions in the automatic evaluation section.

4.2 Human Evaluation

Five experts which are professional translators and interpreters were invited to conduct the human evaluation for 6 primary systems from 6 teams. Each of them was asked to independently rank the translation outputs of 789 sentences. They were required to follow the evaluation guidelines in which the quality of the translations is rated based on two main criteria: Adequacy and Fluency. Adequacy is rated higher than Fluency, however.
Table 3: Automatic evaluation results of the MT task

| Rank | Team       | BLEU   | TER  |
|------|------------|--------|------|
| 1    | EngineMT   | 38.39  | 0.45 |
| 2    | RD-V AIS   | 33.89  | 0.53 |
| 3    | Bluesky    | 32.38  | 0.56 |
|      | Lab-914    | 32.10  | 0.50 |
| 5    | NLP-HUST   | 23.72  | 0.62 |
| 6    | THORLab    | 2.53   | -    |

Table 4: Human evaluation results of the MT task

| Rank | Prize        | Team      | Score  |
|------|--------------|-----------|--------|
| 1    | 1st prize    | Lab-914   | 1.554  |
| 2    | 2nd prize    | EngineMT  | 1.327  |
| 3    | 3rd prize    | RD-V AIS  | 0.864  |
| 4    | -            | Bluesky   | 0.536  |
| 5    | -            | NLP-HUST  | -0.043 |
| 6    | -            | THORLab   | -4.239 |

We used Appraise⁹(Federmann, 2018) - an open-source web-based MT evaluation framework to assist the experts for the evaluation process. Figure 1 is the main interface that the evaluator can rate the outputs of all submitted systems. For each sentence, the evaluator is shown the English source sentence in a context of three sentences: the previous sentence, the currently considered sentence and the followed sentence. Also the golden translation of those three sentences are displayed as the references. The evaluator needs to rank each system’s output from 1 (best) to 6 (worst), and tied ranking is allowed for two or more systems having the same translation quality.

The rankings from 5 experts were converted to pair-wise rankings (number of wins, loses and ties between a pair of two systems). Then they were combined into overall scores using a variant of TrueSkill (Sakaguchi et al., 2014), a sophisticated algorithm considering not only the average number of wins but also how difficult the task is and the variance of each system’s translation quality.

5 Evaluation Results

5.1 Automatic Evaluation

We evaluated all the submissions, including contrastive systems and informed the participants BLEU and TER of their systems. But only the primary systems are ranked, and by their BLEU scores within statistically significant differences ($p \leq 0.05$). The ranking of the teams with corresponding BLEU and TER scores are described in Table 3.

Excepts the team THORLab seemed to have some errors in their submission, other teams produced decent outputs. Unsurprisingly, EngineMT led the board with a considerably large margin to the second team RD-V AIS, maybe because of their domain adaptation techniques. Bluesky and Lab-914 were shared the third rank when the differences between their BLEU scores is not significantly obvious. Notably, based on TER, Lab-914 were ranked second, only after EngineMT.

In some internal test, we realized that other data excepts the OpenSubtitle were high quality and would bring improvements to the systems that use them, even their domain are not news. Lab-914 used a large transformer model and all the provided data, but their BLEU score are not on pair with RD-V AIS which used only the small, in-domain parallel data. We looked into their outputs and their system description as an attempt to explain the possible inconsistency and we discovered that they did not recover casing of their outputs. BLEU is based on the number of overlapping n-grams so that it is more sensitive to upper-cased and capitalized words than TER which is based on the accuracy of individual words. Later, the human evaluation verified our discovery.

5.2 Human Evaluation

As described in Section 4.2, we gathered the ranking of all the systems from 5 experts and produced a unique score for each systems by using the TrueSkill algorithm with the bootstrap resampling at $p$-level of $p \leq 0.05$. Table 4 lists the final ranking of the teams by human evaluation.

While in automatic evaluation, EngineMT is ranked first, here it goes runner-up, after Lab-914. This verifies our assumption about casing recovery. The automatic evaluation metrics do consider casing in their calculation, but the evaluators do not, following their evaluation guidelines.

6 Findings and Future Plans

VLSP 2020 English-Vietnamese News Translation task is the first official MT task hosted by VLSP organizers and it is also the first Vietnamese-related MT evaluation campaign featuring both automatic
and human evaluation. We hope that it would bring scientific and practical values to the VLSP community as well as our society in dealing with the Covid-19 pandemic and in developing useful AI tools.

These are our findings from this VLSP 2020 English-Vietnamese News Translation task:

- English-Vietnamese MT is still a low-resource task with the lack of large-size, high-quality datasets. Back Translation on news data helps improving the overall translation quality. More data, even with mediocre quality and out-of-domain (e.g. OpenSubtitle), when being used to train large models, also brings significantly gains, especially in the human evaluation.

- News might not be a good domain in case we would like to encourage domain adaptation techniques. Monolingual corpora are often crawled from online newspapers and Back Translation might outperform your finest domain adaptation techniques.

- The approaches and techniques are very common and well-known. There is no interesting research finding from the participants.

- There was no submission considering the linguistic characteristics of Vietnamese language or the differences between two languages: English and Vietnamese.

- There were a few participating teams.

We would like to continue hosting MT evaluation tasks in the near future with these plans in mind:

- More language directions in both well-resource and low-resource conditions

- More data in the popular MT tasks

- Consider some useful and interesting domains such as medical, law or technical domains.

- Spread the words to attract more participants working on interesting MT tasks.

7 Acknowledgment

This task is mainly conducted by the Speech and Language Processing Department, VinGroup Big Data Institute with the help of professional translators and interpreters. We would like to thank Vingroup Innovation Foundation (VINIF) for the financial support and the Association for VLSP for organizing support during the campaign.

References

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In Conference of European Association for Machine Translation, pages 261–268.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2017 Evaluation Campaign. In International Workshop on Spoken Language Translation (IWSLT’15), Danang, Vietnam.

Christian Federmann. 2018. Appraise - Evaluation Framework for Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 86–88.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pages 177–180.

Duc Cuong Le and Thi Thu Trang Nguyen. 2020. Vietnamese-English Translation with Transformer and Back Translation in VLSP 2020 Machine Translation Shared Task. In Proceedings of the Sixth Conference of the Association for Vietnamese Language and Speech Processing (VLSP 2020).

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC’2016).

Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. 2013. EVBCorpus - A Multi-layer English-Vietnamese Bilingual Corpus for Studying Tasks in Comparative Linguistics. In Proceedings of the 11th Workshop on Asian Language Resources, pages 1–9.

Thi-Vinh Ngo, Minh-Thuan Nguyen, Hoang-Minh-Cong Nguyen, Hoang-Quan Nguyen, Phuong-Thai Nguyen, and Van-Vinh Nguyen. 2020. The UET-ICTU Submissions to the VLSP 2020 News Translation Task. In Proceedings of the Sixth Conference of the Association for Vietnamese Language and Speech Processing (VLSP 2020).
Ngoc Phuong Pham, Quang Chung Tran, Quang Minh Nguyen, and Hong Quang Nguyen. 2020. A Report on the Neural Machine Translation in VLSP Campaign 2020. In Proceedings of the Sixth Conference of the Association for Vietnamese Language and Speech Processing (VLSP 2020).

Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, et al. 2016. Introduction of the Asian Language Treebank. In 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pages 1–6. IEEE.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 1–11.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In Association for Computational Linguistics (ACL 2016), Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In Association for Computational Linguistics (ACL 2016), Berlin, Germany.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC’2012), pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.