The UET-ICTU Submissions to the VLSP 2020 News Translation Task

Thi-Vinh Ngo[1], **Minh-Thuan Nguyen**[2], **Minh Cong Nguyen Hoang**[2]
Hoang-Quan Nguyen[2], **Phuong-Thai Nguyen**[2], **Van-Vinh Nguyen**[2]

[1]***University of Information and Communication Technology, TNU, Viet Nam***
[2]***University of Engineering and Technology, VNU, Viet Nam***
ntvinh@ictu.edu.vn, npthai@vnu.edu.vn

Abstract

Our UET-ICTU team includes members from the University of Engineering and Technology (UET) and Thai Nguyen University of Information and Communication Technology (ICTU). We participate in the VLSP 2020 Shared Task for Machine Translation which focuses on the news domain translation in one direction English → Vietnamese. Our neural machine translation (NMT) system uses Back Translation (BT) of monolingual data in the target language to augment synthetic training data. Besides, we leverage the Term Frequency and Inverse Document Frequency (TF-IDF) method to data selection close to the in-domain from other monolingual and parallel resources. To enhance the effectiveness of the system translation, we also employ other techniques such as fine-tuning and assembly translation. Our experiments showed that the system can achieve a significant improvement in BLEU score up to + 16.57 overcoming the in-domain baseline system.

1 Introduction

The University of Engineering and Technology (UET) and Thai Nguyen University of Information and Communication Technology (ICTU) participate in the VLSP 2020 Shared Task for Machine Translation on news domain translation from English to Vietnamese (Ha et al., 2020). From datasets in different domains of the Shared Task, we use various strategies to improve the quality of translation in the news domain.

Data selection Data selection techniques help MT systems better translate on a specific domain by eliminating irrelevant data from resources outside the in-domains. This reduces training time but still preserve performance when using smaller datasets instead of training on the large ones. Many works show several methods to select sentences close to background corpus such as: (Axelrod et al., 2011; van der Wees et al., 2017) compute scores for sentences out of domain corpus based on cross-entropy difference (CED) (Moore and Lewis, 2010) from language models; (Wang et al., 2017; Zhang and Xiong, 2018) use sentence embeddings to rank source sentences. This method is only suitable for recurrent networks in NMT. (Wang et al., 2018; Zhang and Xiong, 2018) investigate the translation probability $P(y|x, \theta)$ to be a dynamic criterion to extract sentence pairs during the training process. (Peris et al., 2016) train a neural network classifier to classify sentences into negative or positive fields. These works require training either language models or neural networks and they are less effective in the data sparse situations. (Silva et al., 2018) show empirical results in three various strategies as CED (Moore and Lewis, 2010), TF-IDF (Salton and Yang, 1973) and Feature Decay Algorithms (FDA) (Poncelas et al., 2017). They show that the TF-IDF method has achieved the best improvements in both BLEU and TER (Translation Error Rate) measures. This technique is simple, fast, and does not require training language models or neural networks. Therefore, in this paper, we will leverage it to rank sentences in the scenario that in-domain corpus is small. The detail of this method will be presented in section 3.

Using monolingual resource Monolingual data is used widely in machine translation (MT) (Sennrich et al., 2015; Ha et al., 2017; Lample et al., 2018; Siddhant et al., 2020) due to its widely available. In this paper, we create additional synthetic parallel training data using BT method in (Sennrich et al., 2015) and investigate its effectiveness in our MT systems by combining with genuine parallel data.

Fine-tuning (Luong and Manning, 2015; Zoph et al., 2016) have proposed the fine-tuning process to transfer some of the learned parameters from the parent model to the child model and have

shown significant improvements in many translation tasks. Our systems also fine-tuning on subcorpus (a smaller corpus is extracted from a large corpus) to achieve the best translation effectiveness.

Ensemble translation Ensemble translation (Luong et al., 2015) enable to incorporate the outputs of trained models to enhance translation systems. We attempt to investigate this strategy in our MT system.

Our paper demonstrates a substantial improvement in translating the news domain from the VLSP 2020 Shared Task when combining the aforementioned techniques.

In Section 2, we present an overview of Neural Machine Translation and focus on the transformer architecture. The details of the methods in our paper are presented in Section 3. The settings of the translation system and experimental results are discussed the Section 4. Related works are showed in Section 5. Finally, conclusions and future works are described in Section 6.

2 Neural Machine Translation

Neural Machine Translation (Cho et al., 2014; Sutskever et al., 2014) uses memory units such as Gated Recurrent Units (GRU) or Long Short-Term Memory (LSTM) to overcome the exploding or vanishing gradient problem in recurrent networks. They suggest a new architectural type for MT systems in the form of end-to-end. It includes an encoder to present the sentence in the source language including n tokens $X = (x_1, x_2, ..., x_n)$ into the continue space and a decoder to generate the predicted sentence $Y = (y_1, y_2, ..., y_m)$ in the target language containing m tokens.

The attention mechanism (Luong et al., 2015a; Bahdanau et al., 2015) is considered as the soft-alignment between a source sentence and the corresponding target sentence to enhance the effectiveness of the systems.

Due to the fact that recurrent neural networks (RNN) have limited parallelization in the training process, (Vaswani et al., 2017) propose the transformer architecture that may be highly parallelizable as well as better in translating long sentences. In the transformer, instead of using GRU or LSTM units, a word attends to the other words in a sentence using the self-attention mechanism as the following:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V \quad (1)$$

where K (key), Q (query), V (value) present the hidden states of tokens in the input sentence from encoder or decoder and d is the size of the input.

The attention mechanism in the transformer is the variant of the original attention (Luong et al., 2015a; Bahdanau et al., 2015) when we replace queries by the decoder's hidden states while keys and values come from the encoder's hidden states in the equation 1.

The NMT system is trained to optimize its parameters θ through minimizing the maximum likelihood of all sentence pairs.

$$\mathcal{L}(\theta) = \frac{1}{T}\sum_{k=1}^{k=T} logP(Y^k|X^k;\theta) \quad (2)$$

where T is the number of sentence pairs in the bilingual corpus.

3 The strategies improve our MT system

3.1 Data selection

As mentioned in section 1, in this paper, we utilize the TF-IDF method (Salton and Yang, 1973) to extract a subset of data from large datasets. In the method, TF is the term frequency which presents the ratio between the number of times a term (a word or a sub-word) appears in a sentence and the total number of terms in the sentence. IDF is the inverse document frequency which specifies the ratio between the total number of documents and the number of documents containing the term. Thus, an in-domain corpus D contains T sentence pairs, the TF-IDF score of the token w in the sentence s in the general domain G is evaluated as:

$$score_w = TF - IDF_w = \frac{F_w^G}{W_s^G}.\frac{T^D}{K_w^D} \quad (3)$$

where F_w^G is the frequency of w in s, W_s^G is the length of s, and K_w is the number of sentences in D contain w.

The score of the sentence $s \in G$ is calculated as :

$$score_s = \sum_{i=1}^{i=W_s^G} score_{w_i} \quad (4)$$

These scores are then used to rank sentences in corpus G. The sentence which has the highest

score is nearest to the background corpus, and vice versa.

Our work employs this technique to extract both bilingual and monolingual data.

3.2 Back Translation

In order to improve the translation system from the source language X to the target language Y, (Sennrich et al., 2015) trained the backward translation system from Y X, and it is then used to infer monolingual data from the language Y to predict hypotheses in the language X. We will gain the synthetic bilingual data and it is then mixed with the original bilingual data to augment the training corpus. This technique is called Back Translation (BT).

Our paper applied BT to generate pseudo parallel data English-Vietnamese in the limited bilingual data scenario. In reality, the monolingual data is available but the inference in NMT takes a long time, so we leverage the data selection mentioned in section 3.1 to filter monolingual data.

3.3 Fine-tuning

NMT systems are trained on a large corpus, and then continuously fine-tuned on the in-domain corpus to achieve better performance. We train the NMT system on the mixed datasets from various domains, and then fine-tuning on a smaller corpus extracted from original generic corpus using the strategy in section 3.1.

3.4 Ensemble Translation

The outputs of NMT models can be saturated together to predict better hypotheses. We call this ensemble translation (Luong et al., 2015). The combination vector is simply selected from maximum, or minimum or, average (can be then normalized) probabilities of the output vectors. In this work, we attempt to exhaustive the mean of probabilities from three models and find that a trivial improvement comparing to an individual one.

4 Experiments

4.1 Datasets

Our work only employs the datasets from the VLSP 2020 Shared Task for Machine Translation. It includes six bilingual corpora in divergent domains and one Vietnamese monolingual corpus. This Shared Task focuses on translating the News domain. The bilingual datasets are described in Table 1.

No.	Domains	Training	dev	test
1	News (in-domain)	20K	1007	1220
2	Basic	8.8K	-	-
3	EVBcorpus	45K	-	-
4	TED-like	546K	-	-
5	Wiki-ALT	20K	-	-
6	Open subtitle	3.5M	-	-

Table 1: The English-Vietnamese parallel datasets are used in our work

We use 5 datasets from (1) to (5) for training experiments, the Open subtitle corpus is only used for learning sub-word units in English. The Vietnamese monolingual corpus which includes 20M sentences is exploited for the back translation.

4.2 Preprocessing

We firstly tokenized and true-cased English texts using Moses's scripts. Next, we concated all 6 bilingual corpora to learn 40.000 operators Byte Pair Encoding (BPE) codes like (Sennrich et al., 2016). Lastly, the tokenized and true-cased texts were applied to BPE codes.

Vietnamese texts were tokenized and true-cased using Moses's scripts.

4.3 Systems and Training

We conduct our experiments using the source code from `NMTGMinor`[1]. Our NMT system included four layers for both encoder and decoder and the embedding and hidden sizes are 512. The systems are trained with each mini-batch size of 64 sentence pairs (except the baseline system uses 32 sentence pairs). The vocabulary sizes are 50K tokens for both source and target sides. We use dropout with a probability of 0.2 for embedding and attention layers. The Adam optimizer is applied for updating parameters with an initial learning rate of 1.0. A beam size of 10 is employed for the decoding process.

We train our NMT systems after 50 epochs, and then they are fine-tuned on extracted and in-domain corpus to enhance the accuracy.

4.4 Results

We present empirical results in two measures: BLEU (Papineni et al., 2002) and Translation Er-

[1] https://github.com/quanpn90/NMTGMinor

ror Rate (TER) (Snover et al., 2006). They are implemented in sacreBLEU[2].The higher scores in BLEU specify the better translations while the lower scores in TER indicate better ones. Table 2 shows our experimental results.

In-domain system (baseline) We train the baseline system on News corpus. We learn 10K operators BPE codes and then English texts are applied them.

News + 4 corpus We find that the Open subtitle corpus contains sentences that are not news domain. Therefore, we only combine the background corpus with the 4 remaining corpora. We have shown the improvements of +14.13 BLEU points and -0.259 TER scores.

+ Back Translation We rank sentences from Vietnamese monolingual corpus using the data selection method mentioned in section 3.1, and then extract the top 200K sentences from the ranked text. We employ the backward translation system from Vietnamese → English to generate synthetic bilingual data. The synthetic data are then concatenated to the corpus in the system (2) to train again. We obtain +15.31 BLEU and -0.295 TER points.

+ Fine-tuning on ranked corpus We rank 4 parallel corpora from (2) to (5) in Table 1 using the TF-IDF method in section 3.1 again, and then we also extract the top 200K sentence pairs. The extracted data is combined with the background corpus to continuously fine-tuning the system (3) with an initial learning rate at 0.5. The improvements can be found as +16.21 BLEU and -0.297 TER scores.

+ Fine-tuning on News domain We continue to fine-tune the system (4) with an initial learning rate at 0.25 in the in-domain corpus to gain the best performance, + 16.57 BLEU and -0.3 TER scores.

+ Ensemble translation We combine the output of three best models from the system (5) using the method mentioned in 3.4. We see that our system does not improve.

5 Related Work

NMT systems are restricted in domain translation, therefore, previous works have proposed a variety of data selection techniques to retrieve sentences that are the most related to a specific domain. (Axelrod et al., 2011; van der Wees et al., 2017) leverage language model to estimates the cross-entropy difference (CED) (Moore and Lewis, 2010) for sentences from generic domain. (Wang et al., 2017; Zhang and Xiong, 2018) employed the embedding vectors in the source space from NMT systems to rank sentences. (Wang et al., 2018; Zhang and Xiong, 2018) suggested a dynamic selection based on translation probability to classify sentences during the training process. (Peris et al., 2016) train a neural network to separate sentences into individual domains. These methods are quite complex because they require training neural networks or language models. (Silva et al., 2018) conducted experiments on CED, TF-IDF, FDA, and observe that the TF-IDF strategy is very fast and effective for data selection. In this works, we investigate this method again in the English-Vietnamese translation task.

Due to the lack of bilingual data, some prior studies exploited monolingual data in different ways. (Sennrich et al., 2015) proposed BT method by using used monolingual from the target language. (Ha et al., 2017) shown the mix-source technique to create synthetic data by making a copy of the target language. (Lample et al., 2018) used monolingual data for unsupervised NMT. (Siddhant et al., 2020; Ngo et al., 2020) investigated monolingual data in multilingual NMT. Our work also attempts to using BT method to enhance our NMT system in the data sparse issue.

To gain the best performance in the background domain, (Luong and Manning, 2015; Zoph et al., 2016) demonstrate the effectiveness when transferring the knowledge from the parent model to then child model by the fine-tuning technique. We also apply this approach to our NMT system to achieve better improvements. Besides, we attempt to estimates the quality of the system when using ensemble translation in (Luong and Manning, 2015)

6 Conclusion and Future Work

Our NMT systems have achieved significant improvements when integrating simple techniques such as data section, BT, fine-tuning. In the future, we will leverage more data from other resources as well as using pre-trained models to improve the translation system.

7 Acknowledgments

We would like to thank the organizers and sponsors of the VLSP 2020. We also thank reviewers

[2] https://github.com/mjpost/sacrebleu

No.	Systems	dev		test		offical test	
		BLEU	TER	BLEU	TER	BLEU	TER
1	News corpus (In-domain, baseline)	33.42	0.550	31.66	0.568	21.82	0.753
2	News + 4 corpus (basic + evb + Ted-like + wiki-alt)	46.40	0.427	45.13	0.436	36.12	0.494
3	+ Back Translation	46.35	0.418	45.47	0.436	37.13	0.458
4	+ fine-tuning on ranked bilingual data	48.23	0.399	47.32	0.415	38.03	0.456
5	+ fine-tuning on News corpus	48.94	0.399	48.03	0.405	**38.39**	**0.453**
6	+ Ensembly translation	**49.02**	**0.393**	**48.08**	**0.404**	38.32	**0.453**

Table 2: The results of our English → Vietnamese MT systems are measured in BLEU and TER scores.

who review our paper carefully and give us helpful comments.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2017. Effective Strategies in Zero-Shot Neural Machine Translation.

Thanh-Le Ha, Van-Khanh Tran, and Kim-Anh Nguyen. 2020. Goals, challenges and findings of the vlsp 2020 english-vietnamese news translation shared task. In *VLSP 2020*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Thi-Vinh Ngo, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh, and Le-Minh Nguyen. 2020. Improving multilingual neural machine translation for low-resource languages: French, english - vietnamese. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 55–61.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Álvaro Peris, Mara Chinea-Rios, and Francisco Casacuberta. 2016. Neural networks classifier for data selection in statistical machine translation. *CoRR*, abs/1612.05555.

Alberto Poncelas, Andy Way, and Antonio Toral. 2017. Extending feature decay algorithms using alignment entropy. pages 170–182.

G. Salton and C. S. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation.*, 29(4):351–372.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

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

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Brussels, Belgium. Association for Computational Linguistics.

Matthew Snover, Bonnie J. Dorr, R. Schwartz, and L. Micciulla. 2006. A study of translation edit rate with targeted human annotation.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2018. Dynamic sentence sampling for efficient training of neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 298–304, Melbourne, Australia. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Shiqi Zhang and Deyi Xiong. 2018. Sentence weighting for neural machine translation domain adaptation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3181–3190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.