An Empirical Study on Vietnamese-English Natural Language Inference based on Pretrained Language Models with Data Augmentation*

Dinh-Luan Ngo[1,2], Hieu-Kien Ngo Le[1,2], and Thuy-Ngan Nguyen Luu[1,2]

[1] Vietnam National University, Ho Chi Minh City, Vietnam
[2] University of Information Technology, Ho Chi Minh City, Vietnam
ngannlt@uit.edu.vn
{18521064,18520952}@gm.uit.edu.vn

Abstract. Recently, Natural Language Inference has attracted the attention of research communities due to its application in the Natural Language Processing fields. In this paper, we describe an empirical study of data augmentation techniques with various pre-trained language models on the bilingual dataset which is presented at the VLSP 2021 Vietnamese and English-Vietnamese Textual Entailment. We investigate and compare the effectiveness of a monolingual and multilingual model by applying the machine translation tool to generate new training set from original training data. Our experimental results show that fine-tuning a pre-trained multilingual language XLM-R model with an augmented training set gives the best performance. Our system ranked third at the shared-task VLSP 2021 with about 0.88 in terms of F1-score.

Keywords: Vietnamese and English-Vietnamese Textual Entailment · Pretrained language models · VLSP 2021 dataset · Data Augmentation.

1 Introduction

In recent years, Natural Language Inference (NLI) has attracted the attention of a large number of research communities. It is not only important in academics but also is extremely useful for many information monitoring applications, namely opinion mining, brand and reputation management, and especially fake news system and applications involving semantic understanding [1].

In soving NLI problems, the common approach is to examine the relationship between a pair of sentences or paragraphs (premise and hypothesis) whether they are semantically agree, disagree or neutral to each other [2]. In the shared-task VLSP 2021: "Vietnamese and English-Vietnamese Textual Entailment". This task is presented as a multi-class classification problem involving *sentences_1* and *sentences_2* and the output is a relation of two sentences. Table 1 presents an example in this task.

* Supervisor. Email: ngannlt@uit.edu.vn

Table 1. An example for the task of classifying the "premise" and "hypothesis" pairs. The "premise" can be written in English or Vietnamese, but the "hypothesis" is only written in Vietnamese.

Premise: Vietnamese Tổng thống Trump được cho là đang trải qua các triệu chứng nhẹ của virus corona, bao gồm ho, nghẹt mũi, sốt nhẹ và mệt mỏi. *(President Trump is said to be experiencing mild symptoms of the coronavirus, including cough, stuffy nose, low-grade fever and fatigue).*
Hypothesis: Vietnamese Mặc dù Tổng thống Trump đã dương tính với COVID-19 nhưng vẫn chưa xuất hiện triệu chứng của bệnh. *(Although President Trump has tested positive for COVID-19, he has yet to show any symptoms of the disease).*
Label: Disagree

In Natural Language Processing (NLP), most machine learning models typically depend on the quality and amount of training data; however, collecting and annotating sufficient data is a complicated task. In addition, most available datasets are annotated for rich-resource languages such as English, Chinese, and others. Many studies have focused on data augmentation techniques for the low-resource language to solve this gap. Data augmentation is one of the techniques to increase the number of samples from an existing dataset and enhance the morphological and diversity in the training dataset. Therefore, decreasing dependency on potentially costly and time-consuming data collecting. This technique is simple yet powerful and can work effectively in numerous languages and tasks in NLP.

The concept of back-translation first is applied in the work of [3]. The authors used the back translation method to create more training samples to improve the model's performance. Besides, this technique is more commonly utilized in other tasks such as Sentiment Analysis, Question Answering. On the NLI task, it is more difficult to classify the relation of two sentences because modified versions of the original sentences may no longer have the same meaning and entailment. This paper takes advantage of the peculiar bilingual dataset in the VLSP 2021 competition, presents an empirical study on the sentence pair reversal data augmentation technique. The sentence pair reversal technique translates a sentence from one language to another language. This technique can help our system learn evenly distributed and not focused on a specific language; therefore, our model can learn contextually better than the original dataset. However, the threat is that data may lose meaning during translation, or even worse, or be misleading. As a result, we must exercise caution in terms of accuracy and make excellent use of translation. For that reason, in this paper, we focus on investigating the two available translation techniques and choose the one that provides the best results.

Our study is conducted to try to answer two research questions as follows:

- Consider whether the cross-lingual transfer and automatic translation can perform well in state-of-the-art pre-trained language models such as XLM-R, PhoBERT.

– Whether the sentence pair reversal technique helps us achieve better results or not, and whether it will interfere with the data noise or not.

The organization of the paper is as follows: In Section 2, we will discuss some related works on this topic, and Section 3, we will explain more about our system overview. Section 4 is our results and the performance analysis. Section 5 is the conclusion and the future work.

2 Related Work

Natural Language Inference: Early work on natural language inference has been performed on rather small datasets with more conventional methods [4]. [5] made available the SNLI dataset with 570,000 human-annotated sentence pairs. They also experimented with simple classification models as well as simple neural networks that encode the premise and hypothesis independently. The Multi-Genre Natural Language Inference (MultiNLI) corpus [5] has 433K sentence pairs. Its size and mode of collection are modeled closely like SNLI. MultiNLI offers ten distinct genres of the English language for the task of natural language inference. XNLI [6] is an evaluation set grounded in MultiNLI for cross-lingual understanding (XLU) in 15 different languages that include low-resource languages like Vietnamese. In VLSP 2021: Vietnamese and English-Vietnamese Textual Entailment, we are provided with the VLSP dataset on NLI with 17,200 sentences which is annotated and bilingual of English and Vietnamese.

Data Augmentation: Data augmentation has been proven to be an effective way to tackle challenges sets [7, 8].Various works have demonstrated that by correcting the data distribution, bias can be reduced significantly [9, 10]. Fadaee et al. use contextualized word embeddings to replace target words. They use this text augmentation to validate the machine translation model in [11]. Kobayashi proposed to use a bi-directional language model in [12]. After selecting the target word, the model will predict possible replacement by giving surrounding words. As the target will exist in any position of the sentence, bi-directional architecture is used to learn both rightward and leftward context.

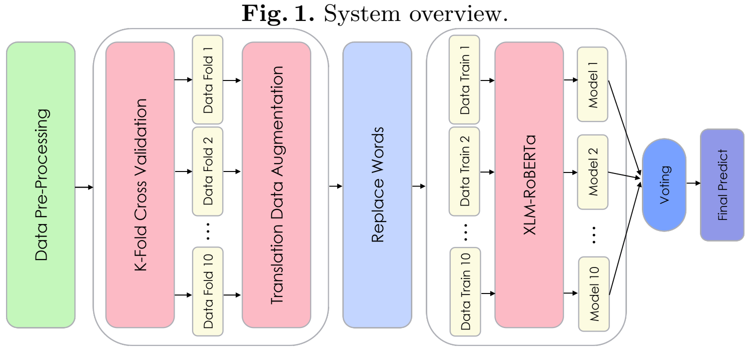
3 System Overview

In this section, we describe our approach to solve this task, including the following sub-sections: 1) Data Pre-processing; 2) Data Augmentation; 3) Classification Architecture; 4) Experiment Setup. Our overall system is shown in Figure 1.

3.1 Data Pre-processing

To extract useful features, we applied different pre-processing steps on the text input, which are outlined below:

Fig. 1. System overview.



- **Step 1**: We removed characters such as punctuation, icon, hashtag, link URL, or words that are not alphanumeric in two sentences.
- **Step 2**: Removing the null and noise samples in the training set (usually containing the only character "bở"). This must be a small mistake in the training set.
- **Step 3**: After that, we replace words with synonyms without affecting the meaning of the sentence based on the manually dictionary from the training set. For example, same words such as "Coronavirus", "COVID-19", "SARS-COV-2" were replaced to "corona".

Besides, we applied the removing "stop words" technique in our pre-processing steps; however, the results were ineffective. Removing stop words in this task might break the link between the "premise" and the "hypothesis" sentences, resulting in unsatisfactory results. Table 2 shows the statistic after applying pre-processing steps on both training and testing datasets.

Table 2. Summary of the dataset after applying the pre-processing steps.

	Vi-Vi	**En-Vi**	**Total**
Training set	8606	7500	16177
Testing set	2118	2059	4177

3.2 Data Augmentation

Because of the advancement of machine translation models, data augmentation has grown in popularity in recent years. There are some available machine translation models to translate between Vietnamese and English language such as the Google Cloud Translation API[3] and the VietAI Machine Translation[4]. From a

[3] https://cloud.google.com/translate

[4] https://github.com/vietai/SAT

Table 3. Examples of sentence pairs reversal data with P as a Premise and H as a Hypothesis.

Original pairs	Augmented pairs
P1(en): One of the few silver linings of the novel corona virus is that it mostly spares kids. H1(vi): Tất cả mọi người đều có khả năng lây nhiễm vi-rút corona như nhau, đặc biệt là trẻ nhỏ.	P1'(vi): Một trong số ít những điều đáng chú ý của corona virus mới là nó hầu như không để lại cho trẻ em. H1'(en): Everyone has the ability to infect Corona viruses equally, especially young children.
P2(vi): Theo Sở Y tế Bang Hawaii, hiện đã có 607 trường hợp được xác định nghi nhiễm Covid-19 ở đây. H2(vi): Hawaii là bang duy nhất chưa ghi nhận ca nhiễm COVID-19 nào.	P2'(en): According to the Hawaii Department of Health, there are currently 607 cases of Covid-19 identified cases here. H2'(en): Hawaii is the only state that has not recorded Covid-19.

limited training data source, it will automatically generate more training data and is considered semi-supervised learning [3, 13].

After experimenting with the paid version of Google Translation API and free version of VietAI Machine Translation, we found that the model translated by the Google Translation API give better results. Therefore, we used Google Translation API as the main translation tool for our experiments. There are three strategies based on the machine translation tool in our paper as follows:

- **Sentence Pairs Reversal:** Given a source and target sentence pair (P,H). We would like to change it such that the semantic equivalence between P and H is preserved while the training instances are as diverse as feasible. Basically, this approach aims to create new sentences by reversing the pair of sentences (P,H) into (P',H'). In this way, we can increase the training samples for our model.
- **Convert to English:** Based on our survey, most pre-trained language models were developed for the English language, therefore, we translate whole Vietnamese sentences to English and experiment on fine-tuning pre-trained language models such as XLM-R and Albert [14]. The sentences in the test set are also translated to English for the evaluation process.
- **Convert to Vietnamese:** As similar, we convert whole English sentences to Vietnamese sentence on the training and testing set, then train them by using the PhoBERT [15] and XLM-R model.

Table 3 shows examples of data that have been translated with Google Cloud Translation API. Table 4 describes our dataset after applying Tranlation Data Augmentation. We are provided with a training dataset from the organizers (VLSP dataset). We used GG translation API to translate the dataset to English and named it VLSP_en. The VLSP_en will be used to evaluate on the English private test data by ALBERT and XLM-R models. Following, the origi-

Table 4. Summary of the dataset after using data augmentation method.

Training Data	Original Data	Data Augmented	Data Training
VLSP	16 177	-	16 177
VLSP_en	16 177	16 177	16 177
VLSP_vi	16 177	7500	16 177
VLSP_au	16 177	16 177	32 354
VLSP_au+en	16 177	23 676	39 853
VLSP_au+vi	16 177	23 676	39 853

nal dataset already contains 8,676 vi-vi sentences, so we translated the remaining 7,500 en-vi sentences to Vietnamese and named it VLSP_vi. The VLSP_vi will be used to evaluate on the translated private test data by PhoBERT. We also continue to translate the data by paired sentences reversal method in Section 3.2 and named it as VLSP_au (including original dataset). And the final dataset is a combination of the VLSP_en, VLSP_vi with VLSP_au tuples we named it as VLSP_au+en and VLSP_au+vi to evaluate efficiency of multilingual tranfer learning technique.

3.3 Classifier Architecture

One of the purposes in our paper is to investigate the performance of multilingual models on bilingual dataset and monolingual models on the translated dataset. All mentioned models were used in the base and large version, except for ALBERT.

Multilingual model: We chose XLM-R over mT5 [16] and mBERT [17] because XLM-R generally performs better than mT5, mBERT at the same model size (see original paper for details). The work of [18] demonstrated that the XLM-R model is currently the best multilingual model for Vietnamese language.

Vietnamese Monolingual model: PhoBERT is one of the best monolingual models for various tasks in the Vietnamese NLP topic. To employ this model, we use the VnCoreNLP [19] to perform word and sentence segmentation on the input as to their recommendation.

English Monolingual model: As above mentioned, we use two pre-trained language models such as XML-R and Albert to train model on whole translated English dataset.

Experiment Setup: To choose our best model, we ran various experiments to test the effectiveness of the different approaches. All experiments have been carried out with a learning rate set at 1e-5, using Adam optimizer. The batch size is selected in a set of {4, 8,16} and 16 is the best value in our experiments. With maximum sentence length, we used {37,64,100,111,128} where 37 is the average length of dataset and 111 is the maximum length. We found that with a maximum sentence length at 100 we got the best results and did the training in 3 epochs.

Table 5. The results f1-score of the XLM-R model on each data set.

Fold	**VLSP**	**VLSP en**	**VLSP au**	**VLSP au+en**
1	85.22	85.10	88.33	88.60
2	84.67	84.90	87.66	87.10
3	86.55	84.66	**89.33**	86.33
4	**87.43**	86.33	86.90	85.80
5	85.10	85.60	87.90	85.20
6	86.66	84.90	86.22	86.00
7	86.33	85.33	86.66	87.55
8	87.10	86.83	88.20	**89.10**
9	85.33	**87.20**	88.33	88.33
10	84.92	84.66	84.40	87.90
average	85.93	85.55	**87.79**	87.20

At VLSP 2021, we formulate our training data in a 10-fold cross-validation manner. From the models, we obtain the average probability of the response prediction. Then, we use ensemble methods as hard voting to make the final evaluation on the private test set. Table 5 shows our results when training the model with the above data sets using the same parameters. However, because we were given a private test set in the past study, we only trained on training data and assessed it on private test set.

4 Results and Analysis

We visualize stopword-removed data using Word Cloud Representation on English in Figure 2 and Vietnamese in Figure 3. In the visualization, we easily notice that words which are semantically similar tend to appear more such as "corona virus", "covid", "virus". By replacing those similar words with one synonym resulted in improvement of model performance, which was also proven by our paper in the VLSP 2021 shared task.

Fig. 2. Visualize with Word Cloud Representation on English dataset.

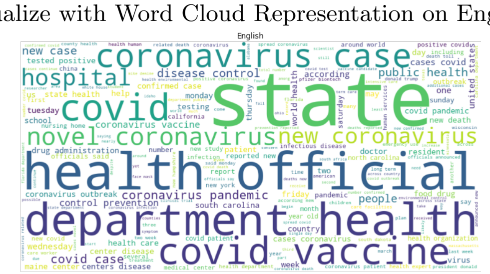


Fig. 3. Visualize with Word Cloud Representation on Vietnamese dataset.

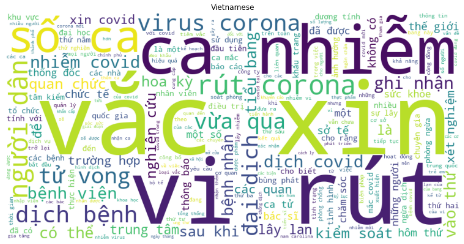


Table 6. The experimental results of various Vietnamese and English Monolingual models.

Model	Fine-tuned on	F1-score	Transfer
ALBERT base	VLSP_en	81.43	mono
ALBERT large	VLSP_en	87.78	mono
PhoBERT base	VLSP_vi	80.02	mono
PhoBERT large	VLSP_vi	86.45	mono
XLM-R base	VLSP_en	80.51	mono
XLM-R large	VLSP_en	**88.07**	mono
XLM-R base	VLSP_vi	81.00	mono
XLM-R Large	VLSP_vi	87.19	mono

Table 7. The experimental results of various Multilingual models with mixed Multilingual and Monolingual.

Model	Fine-tuned on	F1-score			Average	Transfer
		Agree	Disagree	Neutral		
XLM-R large	VLSP original	87.60	84.22	85.96	85.93	Multilingual
XLM-R large	VLSP_au	**90.97**	**88.57**	**90.86**	**90.17**	Multilingual
XLM-R large	VLSP_au+en	89.31	85.96	89.33	88.20	Mixed
XLM-R large	VLSP_au+vi	89.21	85.02	89.82	88.02	Mixed

The main results of monolingual on English and Vietnamese datasets are shown in Table 6 we present the whole model of various sizes. Experiments using models ALBERT, RoBERTa, and PhoBERT on Monolingual data results indicated that they performed worse than the Multilingual model XLM-R on Monolingual data. The reason might be the quality of the machine translation model to translate data to the source language. For VLSP_en and VLSP_vi data, the XLM-R model gives better results than the monolingual model about 1% in terms of F1-score (87%∼ 88%).

Table 7 shows the results of train VLSP_au , VLSP_vi, VLSP_en, VLSP original datasets on XLM-R model. It can be seen that combining the datasets VLSP_en and VLSP_vi reduces the performance of the model. This is partly due to the fact that we employ automatic translation to translate the original VLSP data into VLSP_en and VLSP_vi, which have not been thoroughly tested by experts. The sentence pairs reversal approach improves the VLSP_au data with better results. It shows that this data augmentation technique is well suited to the problem of bilingual data.

Fig. 4. Confusion matrix of XLM-R model fine-tuned on VLSP_au.

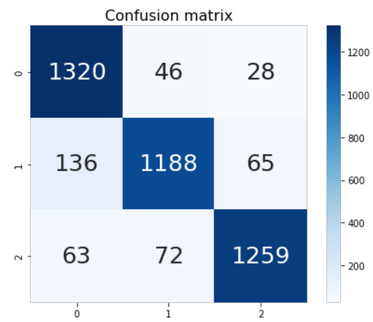


Figure 4 displays a confusion matrix of the best XLM-R model on the VLSP_au dataset. This model is trained on the VLSP_en and VLSP_vi dataset that gives more wrong predictions on the disagree class than the VLSP_au dataset. Therefore, our model is trained on the VLSP_au dataset produces high performances on three classes. The F1-score is more than 90%. This suggests that our model is highly compatible with providing a dataset in VLSP 2021.

5 Conclusion and Future Work

This research presents an empirical study on data augmentation techniques by using Google Cloud Translation API and fine-tuning pre-trained language models. Our experimental results indicated that multilingual models such as XLM_R are suitable for the bilingual NLI dataset. Besides, with the sentence pairs reversal as the data augmentation technique, the performance can be better than other methods about 1-2% in terms of F1-score. For future work, investigating the attention model to extract emphasizing words in sentences that have the "disagree" label might be a new potential research direction.

References

1. N. Reimers, I. Gurevych, Sentence-bert: Sentenceembeddings using siamese bert-networks, in:Proceedings of the 2019 Conference on EmpiricalMethods in Natural Language Processing andthe 9th International Joint Conference on NaturalLanguage Processing (EMNLP-IJCNLP), 2019,pp. 3982–3992.
2. W. Yin, D. Radev, C. Xiong, Docnli: A large-scale dataset for document-level natural language inference, arXiv preprint arXiv:2106.09449.
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)

4. Bill MacCartney. 2009. Natural language inference. Stanford University.
5. Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. pages 632–642.
6. Alexis Conneau, Ruty Rinott, Guillaume Lample, Ad- ina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross- lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natu- ral Language Processing, pages 2475–2485.
7. Jacob Andreas. 2020. Good-enough compositional data augmentation. pages 7556–7566, Online. Associa- tion for Computational Linguistics.
8. Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmen- tation increases robustness to inference heuristics. pages 2339–2352, Online. Association for Computa- tional Linguistics..
9. Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. pages 650–655, Melbourne, Australia. Association for Computational Linguistics.
10. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Or- donez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. pages 15–20.
11. Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. pages 567–573.
12. Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic re- lations. New Orleans, Louisiana. Association for Computational Linguistics.
13. A. Sugiyama, N. Yoshinaga, Data augmentation using back-translation for context-aware neural machine translation, DiscoMT 2019 (2019) 35.
14. W. Y. Wang, D. Yang, That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using petpeeve tweets, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2557–2563.
15. M. Fadaee, A. Bisazza, C. Monz, Data augmentation for low-resource neural machine translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017, pp. 567–573.
16. K. Kafle, M. Yousefhussien, C. Kanan, Data augmentation for visual question answering, in: Proceedings of the 10th International Conference on Natural Language Generation, 2017, pp. 198–202.
17. Z. Lan, M. Chen, S. Goodman, K. Gimpel,P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942.
18. D. Q. Nguyen, A. T. Nguyen, Phobert: Pre-trained language models for vietnamese, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, 2020, pp. 1037–1042.
19. L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498.