VNU Journal of Science: Comp. Science & Com. Eng, Vol. 38, No. 2 (2022) 19-27

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

# vnNLI - VLSP 2021: Vietnamese and English-Vietnamese Textual Entailment Based on Pre-trained Multilingual Language Models

Hoang Xuan Vu, Nguyen Van Tai, Phan Thi Kim Khoa, Dang Van Thin, Duong Ngoc Hao, Nguyen Luu Thuy Ngan*

*University of Information Technology, Vietnam National University, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City*

Received 28 December 2021
Revised 28 March 2022; Accepted 5 May 2022

**Abstract:** Natural Language Inference (NLI) is a high-level semantic task in Natural Language Processing - NLP, and it extends further challenges in the cross-lingual scenario. In recent years, pre-trained multilingual language models (e.g., mBERT-XLM-R, InfoXLM) have greatly contributed to the success of dealing with these challenges. Based on the motivation behind these achievements, this paper describes our approach based on fine-tuning pretrained multilingual language models (XLM-R, InfoXLM) to tackle the shared task "Vietnamese and EnglishVietnamese Textual Entailment" at the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 20211). We investigate other techniques to improve the performance of our work: Cross-validation, Pseudo-labeling (PL), Learning rate adjustment, and POS tagging. All experimental results demonstrated that our approach based on the InfoXLM model achieved competitive results, ranking 2nd for the task evaluation in VLSP 2021 with 0.89 in terms of F1-score on the private test set.

*Keywords:* Vietnamese and English-Vietnamese Textual Entailment, Cross-lingual textual entailment, Pre-trained Multilingual Language Models, Data augmentation, Vietnamese language, VLSP 2021.

## 1. Introduction

The past decades have witnessed the great rise of Artificial Intelligence (AI). One of the central topics in AI is Natural Language Understanding where Natural Language Inference plays an important role, which was pointed out in [1]. Also, in recent years, Natural Language Inference has been used in several NLP applications like Question Answering, Evaluation of Machine Translation systems [2], Fake Information Detection [3], and

* Corresponding author.
*E-mail address:* ngannlt@uit.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.329

Summarization [4], etc. Identifying entailment (agree), contradiction (disagree), or undetermined (neutral) between two sentences (called a "hypothesis" and a "premise") is known asNatural language inference (NLI) or Textual Entailment (TE). NLI was defined as a task of determining whether a natural language hypothesis h can be inferred from a given premise p [5]. The examples are depicted in Table 1.

Table 1. Examples of relations between a premise anda hypothesis: E (Entailment), C (Contracdition), N (Neutral)

| Premise | Hypothesis | Label |
|---|---|---|
| Two girls arewalking on the street. | Some women are walking in a race. | C |
| | They are going for a walk on the street. | E |
| | Two girls are walking with their dogs on the street. | N |

The difference between the VLSP2021's dataset and others are the input sentences, which are in English or Vietnamese and may not be in the same language. For example:

• The Premise: "Researchers in Finland made a computer-simulated model of how a cough could spread particles in a grocery store".

• The Hypothesis: "Một mô hình đã thể hiện được cách thức virus lây lan qua cơn ho trong cửa tiệm tạp hóa".

Because of being one of the new difficulties suggested for this NLI problem, it is the input sentence that complicates the task. The lack of a large and diverse volume of datasets for this challenge has become the key restriction of research development in this line [6]. Besides, the linguistic (dis)similarity between the languages affects machine learning models to extract the information appropriately. To solve this challenge, we consider it as a Cross-lingual natural language inference task [6, 7, 8].

Over the past few years, the potential solution for Cross-lingual tasks is to use pre-trained multilingual language transformer models, such as XLM-R [9], InfoXLM [10], and others [11, 12]. These models are so effective that they have been shown to be a key solution in several NLP tasks [13, 14, 15, 16]. A prominent feature of these models is used for cross-lingual, which can be fine-tuned on a particular task in available annotated datasets called source language, then adapted to the same task in target languages.

Because of the potential of pre-trained multilingual language transformer models, in our work, we present an approach relying on pre-trained multilingual language models consisting of XLM-R and InfoXLM. In addition, we focus on data preprocessing know as POS tagging and some techniques including Cross-validation, Pseudo-labeling (PL), and Learning rate adjustment to address the challenge "Vietnamese and English-Vietnamese Textual Entailment".

In this work, our main contributions can be summarized as follows:

• We conduct an investigation into the benefit of using two of state-of-the-art pre-trained multilingual language models (XLM-R and InfoXLM) to evaluate cross-lingual natural language inference task in VLSP 2021 [17].

• We employ different potential techniques to improve the performance of our system. Moreover, we point out InforXLM model is better than XLM-R in this task.

## 2. Related Work

Recently, NLI datasets [8, 18, 19] were used widely and throughout the development stages of the Cross-lingual NLI problem. Therefore, there have been several researches on the task for improving accuracy on various NLI datasets.

Massively Multilingual Transformers (XLM-R [9], InfoXLM [10], mBERT [11], XLM [20], mT5 [21]) have been shown to have remarkable transfer skills in zero-shot settings. In 2021, [15] investigated the cross-lingual transfer abilities of XLM-R for Chinese and English natural language inference (NLI), with a focus on the recent large-scale Chinese dataset OCNLI [22]. However, the results demonstrated that cross-lingual models often perform well when models are trained on a mixture of English and high-quality monolingual NLI data (OCNLI) and are often hindered by automatically translated

resources (XNLI-zh [8]). Besides, [23] incorporated syntax into natural language inference (NLI) models by using contextual token-level vector representations from a pretrained dependency parser. Like other contextual embedders, their method is broadly applicable to any neural model. Contrasting with the previous models that used complex network architectures, [24] demonstrated a carefully designing sequential inference models based on chain LSTMs can outperform all previous models. In [25], this paper discussed Cutting Edge research on NLI, including recent advance on dataset development, Cutting Edge deep learning models, and highlights from recent research on using NLI to understand capabilities and limits of deep learning models for language understanding and reasoning. In [26], the authors investigated the effectiveness of language modeling, data augmentation, translation, and architectural approaches to address the code-mixed, conversational, and low-resource dataset. Meanwhile, [6] provided a deep neural framework for cross-lingual textual entailment involving English and Hindi. As there are no large datasets available for this task, the authors created their datasets by translating the premises and hypotheses pairs of Stanford Natural Language Inference (SNLI [18]) dataset into Hindi.

Furthermore, during the expanding NLP period, a number of pretrained multilingual language models have been released, which influenced the methodologies in solving the NLI problem. Especially, the performance of these models on NLI datasets was fairly good. Therefore, we utilize some pretrained multilingual language models.

Besides relying on pre-trained multilingual language models, we attempt to use the Psuedo-Labeling method [27] to improve the model's efficiency and focus on data preprocessing. Our approaches are performed on the dataset of VLSP 2021.

## 3. System Overview

### *3.1. Preprocessing and Analysis*

The training set consists of 16200 pairs of sentences and the testing set consists of 4177 pairs of sentences. The detail and analysis is shown in Table 2 and Table 3.

From statistics of the data, we found that some noisy pairs of sentences which have the same content, but different labels. Table 4 demonstrates an explicit example of these pairs of sentences. In particular, we detected 27 pairs of noisy sentences affecting the performance of the models due to their content and labels. To address the problem, we removed these pairs of sentences from the training set. Besides, many synonyms were detected in the training sentences.

Table 2. Statistics on the number of sentences in each language. Lang_1 is the number of Vietnamese sentences. Lang_2 is the number of English sentences. Total is the total number of pairs of sentences in thedataset.

| | Sentence_1 | | Sentence_2 | | Total |
|---|---|---|---|---|---|
| | Lang_1 | Lang_2 | Lang_1 | Lang_2 | |
| Training set | 7503 | 8697 | 0 | 16200 | 16200 |
| Testing set | 1375 | 2802 | 0 | 4177 | 4177 |

Table 3. Summary statistics of training set and testingset. Length is the average sentence length, Vocab_en isthe size of English vocabulary. Vocab_vi is the size ofVietnamese vocabulary.

| | N.o classes | | | Length | Vocab_en | Vocab_vi |
|---|---|---|---|---|---|---|
| | Agree | Disagree | Neutral | | | |
| Trainning set | 5400 | 5400 | 5400 | 30.3 | 8300 | 6981 |
| Testing set | 1394 | 1394 | 1389 | 31.9 | 3937 | 5934 |

Therefore, we processed these words in a standardized format. For example: we converted these words such as “virus corona”, “vi rút”, and “covi” to “Covid-19”. Moreover, through our observation, we noticed that comparative words played a significant role in the output. For that reason, we use a technique named POS tagging in which we emphasized that words in the data by appending < s > tokens before and after these words, then concatenated them to the end of the sentence. Table 5 is an example of the data before and after preprocessing.

Table 4. Examples of pairs of noisy sentences.

| | |
|---|---|
| "id": train_523 | "id": "train_8233" |
| "lang_1": "vi" | "lang_1": "vi" |
| "lang_2": "vi" | "lang_2": "vi" |
| "sentence_1": "bỏ" | "sentence_1": "bỏ" |
| "sentence_2": "bỏ" | "sentence_2": "bỏ" |
| "label": "disagree" | "label": "neutral" |

### *3.2. Approach*

In our work, we utilize two pre-trained multilingual language models: XLM-R and InfoXLM to solve the shared task “Vietnamese and EnglishVietnamese Textual Entailment” in the VLSP2021. The XLM-R model is trained on the CommonCrawl data in 100 languages, while the InfoXLM model is trained on CCNet corpus with 94 languages. Both models include the languages in the task. These two models are used for the following reasons:

• XLM-R: The dataset includes 2 languages, English and Vietnamese. The amount of Vietnamese data makes up the majority of the dataset. For Vietnamese language, XLM-R is the best multilingual model based on the previous research [28].

• InforXLM: The author presented and evaluated this model on the cross-lingual XNLI. They reported test accuracy in 15 languages and showed that the performance of the InfoXLM model gave good results in the specific languages of English and Vietnamese. When fine-tuning the multilingual model on all training sets, the InfoXLM model gave the highest accuracy results, 86.5% for English and 81.0% for Vietnamese.

Table 5. Example of emphasizing comparative words.

| Original Sentence | Pre-processing sentence |
|---|---|
| "Đại dịch virus corona đã giết chết hơn 150.000 người Mỹ, làm gián đoạn nền kinh tế Mỹ và khiến xã hội rơi vào tình trạng hỗn loạnkhi bước vào mùa thu." | "Đại dịch COVID-19 đã giết chết hơn 150.000 người Mỹ, làm gián đoạn nền kinh tế Mỹ vàkhiến xã hội rơi vào tình trạng hỗn loạn khi bước vào mùa thu <S>hơn 150.000 <S>" |

As depicted in the Table 1, the input is a given pair of sentences: premise and hypothesis; the output is a label that determines whether a natural language hypothesis can be inferred from a given premise, such as agree (E), disagree (C), or neutral (N). Therefore, this task can be treated as a multi-class classification problem. To tackle this task, we employ a fine-tuning approach based on the pre-trained language models for this task. After that, we extract the representation of [CLS] token in the last layer as the input representation. This representation is fed directly into the fully connected layer with Softmax activation, which predicts a probability of class for the label. The overview of our approach is displayed in Figure 1.

Besides, we investigated effective machine learning techniques to improve the model’s performance as follows:

• Cross-validation: Because VLSP only provided a training set, we needed to apply the K-fold Cross-Validation (KCV) technique with K=10 on the training set to objectively evaluate the models. Specifically, the training set was divided into 10 folds, which was trained in turn. Then, in every turn, we chose randomly one fold for model to test. To generate more robust

findings, we averaged the model prediction on each fold.

• Pseudo-labeling: Pseudo-labeling [29] is an effective semi-supervised learning method to utilize the abundant unlabeled data via their pseudo labels. In this work, we used training datasets and testing datasets to test. Specifically, we concatenated training datasets and testing datasets which were predicted previously. Then, we predicted testing datasets again to get the final result.

• Learning rate adjustment: As for the search range of each hyperparameter, the learning rates was selected from {1e-5, 2e-5, 5e-6}. We employed experiments progressively and find the best learning rate for each model on training set with Cross-validation technique.

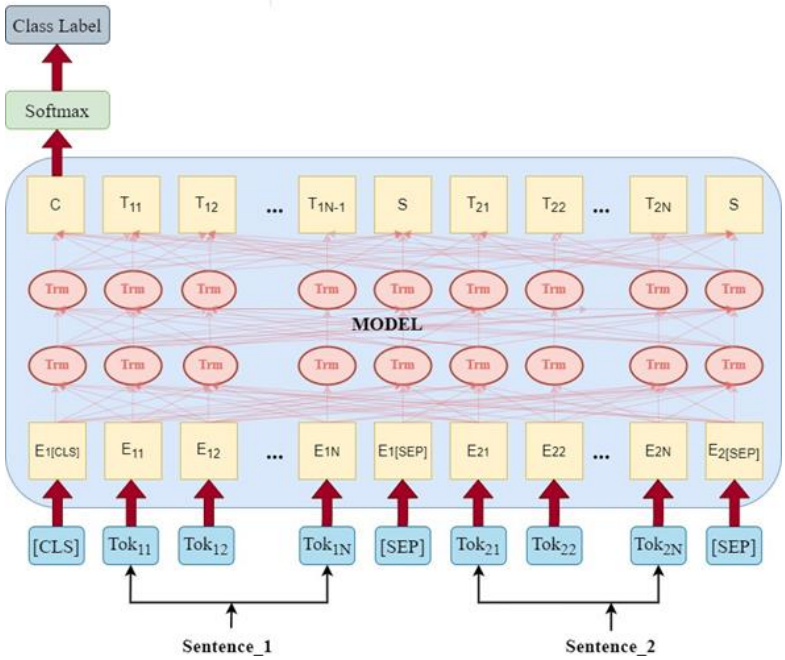


Figure 1. Overall architecture based on fine-tuning pre-trained language models.

## 4. Experiment

### *4.1. Experimental Setting*

Following the given evaluation metrics, in all our experiments, we report the micro F1-score based on evaluation scripts from the task organizing committee.

Table 6. Compare results with other teams on the private test set at shared-task VLSP 2021

| Rank | F1-score |
|---|---|
| First Team | 0.90 |
| Second Team (Ours) | **0.89** |
| Third Team | 0.88 |

As described in Section 3.2, our approach depends on pre-trained language models such as XLM-R and InfoXLM model. We use two base models downloaded from the Hugging Face library [30]. The network's parameters are optimized using the AdamW [31] and a linear learning rate scheduler, which are suggested by the Hugging Face default setup. The hyperparameters that we tune include the number of epochs, batch size, and learning rate. In particular, we use 10 epochs, batch size of 8 for both models. For the XLM-R model, we set learning rate 1e-5. For InfoXLM model, the learning rate is 5e-6. All experiments in this paper are conducted on Google Colab Pro (CPU: Intel(R) Xeon(R) CPU @ 2.30GHz; RAM: 25.51 GB; GPU: Tesla P100-PCIE-16 GB with CUDA 10.1).

*4.2. Result and Discussion*

In this section, we present our experimental results based on two pre-trained multilingual models and our strategies in this task. Firstly, we examine the performance of two base pre-trained language models (InfoXLMbase, XLM-Rbase). As shown in Table 7, we observe that the InfoXLMbase yields better performance than XLM-Rbase by 2.58% on F1-score. To obtain higher scores, we apply different strategies on the models, including Pseudo Labeling (PL), or POS tagging (PT) which is a preprocessing technique.

Table 7. Results of base models on testing dataset under different strategies. The abbreviation is defined as follows: F1-score, PL: Pseudo-labeling, PT: POS tagging

| Type | Model | Precision | Recall | F1-score |
|---|---|---|---|---|
| XLM-R | XLM-R-base | 0.8218 | 0.8227 | 0.8218 |
| | XLM-R-base +PT | 0.8127 | 0.8152 | 0.8126 |
| | XLM-R-base + PL | 0.8220 | 0.8229 | 0.8220 |
| | XLM-R-base + PT + PL | 0.8127 | 0.8152 | 0.8126 |
| InforXLM | InfoXLM-base | 0.8475 | 0.8488 | 0.8476 |
| | InfoXLM-base + PT | 0.8446 | 0.8461 | 0.8445 |
| | InfoXLM-base + PL | 0.8489 | 0.8510 | 0.8489 |
| | InfoXLM-base + PT + PL | **0.8494** | **0.8511** | **0.8493** |

that emphasizes comparative words in the data. However, when using POS tagging (PT), the performance of XLM-Rbase + PT is 0.92% lower than that of XLM-Rbase and InfoXLMbase + PT also decrease 0.32% compared to InfoXLMbase. Besides, we use the Pseudo-labeling technique on two base models and the results have a positive adjustment. The performance of XLM-Rbase + PL is 0.02% higher than that of XLM-Rbase and InfoXLMbase + PL also increase 0.13% compared InfoXLMbase. Although the change of utilizing the Pseudo-labeling approach is small, it demonstrates that this strategy is effective. We combine two techniques: Pseudo-labeling and POS tagging. XLM-Rbase + PL + PT gives a decrease of 0.92%, while infoXLMbase + PL + PT produces the greatest results across all base models.

Although these techniques show potential to experiment on the base model, we also employ these techniques on the large models aiming for achieving the best result to submit the shared task in VLSP 2021. As described in Table 8, this approach achieves significant scores that help us got 2nd place in the competition, as shown in Table 6. In particular, the large models XLMRlarge increase 6.59% to 7.36% compared to XLMRbase with different strategies.With InfoXLMlarge When applying Pseudo-labeling and POS tagging techniques, the model obtained the best results of the investigation models with 0.8493 F1-score.

Table 8. Results of large models on testing dataset under different strategies. The abbreviation is defined as follows: F1-score, PL: Pseudo-labeling, PT: Postagging

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| XLM-R-large | 0.8874 | 0.8874 | 0.8869 |
| XLM-R-large + PL | 0.8867 | 0.8867 | 0.8862 |
| XLM-R-large + PT + PL | 0.8881 | 0.8891 | 0.8879 |
| InfoXLM-large | 0.8921 | 0.8921 | 0.8921 |
| InfoXLM-large + PL | 0.8927 | 0.8930 | 0.8926 |
| InfoXLM-large + PT + PL | **0.8965** | **0.8970** | **0.8964** |

Through our observation from experiments on both base and large models, when using the pseudo-labeling technique, there is small change in the performance of the two models. The performance is unstable when combining POS

tagging and the Pseudo-labeling technique. We realized that the reason the Pseudo-labeling technique didn't work was probably that the ratio between fake and real labels didn't match. This reason is also demonstrated in the work "Realistic Evaluation of Semi-Supervised Learning Algorithms" [29]. It is also affected by the pre-training model. In this case, Pseudo-labeling works fine for the infoXLM model, but degrades the XLMR model's performance. Figure 2 show confusion matrix of model infoXLMlarge + PT and confusion matrix of model InfoXLMlarge + PT: POS tagging + PL: Pseudo-labeling. Although the two best models have no significant change when applying the pseudo-labeling technique, this is also a method that can be used to improve model performance.

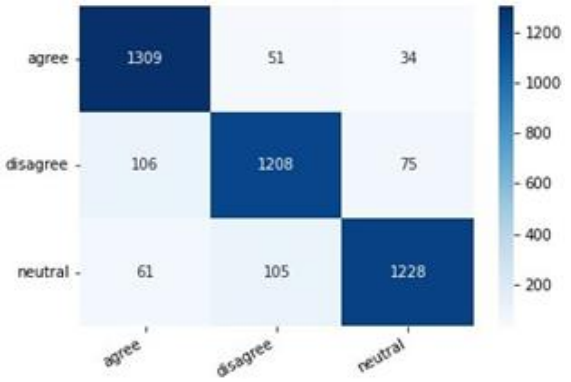


Figure 2. Confusion matrix of model InfoXLM$_{large}$ + PT: POS tagging + PL: Pseudo-labeling.

When applying the POS tagging which emphasizes comparative words in the data, we observe that we found that both XLM-R and InfoXLM give good results and increase model performance.This method has model improvement when applied to XLM-Rlarge and infoXLMlarge, but gives unstable performance on the basis. This method depends on the pre-trained model and the dataset.

After that, we filter out the wrong prediction sentences when applying the comparative word emphasis technique for analysis. We find that a few sentences correctly predicted when applying this technique. Besides, it causes confusion when we predict the pair of sentences which are similar to the example in table 9. When sentence_1 emphasizes "than 800" and sentence_2 emphasizes "hơn 800" the model may predict "agree" instead of the correct label, "neutral". Therefore, to improve the model by this method, the dataset needs to be suitable.

Table 9. Example of incorrect prediction when using comparative word emphasis technique.

{
"id": "test_924",
"sentence_1": "More than 800 nursesat the University of Illinois Hospital in Chicago will go on strike Saturday morning after contract negotiations broke down over nurse-to-patient ratios.",
"sentence_2": "Bệnh viện Đại học Illinois vừa được mệnh danh là bệnhviện lớn nhất của Chicago với tổng số hơn 800 nhân viên y tế và số lượng giường bệnh đạt tới con số 1000.",
"label": "neutral"
}

## 5. Conclusion and Future Works

In this work, we presented our approaches ranked 2nd in VLSP 2021 in solving Vietnamese and English-Vietnamese Textual Entailment based on Pre-trained Multilingual Language Models. In addition, we also compared the performance of two models (XMLR and InforXLM) with different techniques such as cross-validation, pseudo labeling, learning rate adjustment, and POS tagging. From the experimental results, we found that fine-tuning on the InforXLM model obtain better results than that on XLMR.

In the future, we might use the intermediate layers of the InforXLM model to take advantage of the model's huge potential, and we can apply some loss functions to improve the outcome of this task.

## Acknowledgments

Authors would like to thank VLSP 2021 organizers for organizing and providing the NLI

dataset for this research. We also would like to thank the anonymous reviewers for their valuable comments.

## References

[1] B. MacCartney, C. D. Manning, Modeling Semantic Containment and Exclusion in Natural Language Inference, in: Proceedings of the22nd International Conference on Computational Linguistics (Coling 2008), 2008, pp. 521–528.

[2] A. Poliak, Y. Belinkov, J. Glass, B. V. Durme,On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference, arXiv preprint arXiv:1804.09779.

[3] K. Sabarmathi, K. Gowthami, S. S. Kumar, Fake News Detection Using Machine Learning and Natural Language Inference (NLI), in: IOP Conference Series: Materials Science and Engineering, Vol. 1084, 2021, p. 012018.

[4] T. Falke, L. F. Ribeiro, P. A. Utama, I. D. Gurevych, Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2214–2220.

[5] B. MacCartney, Natural Language Inference, Stanford University, 2009.

[6] T. Saikh, A. De, D. Bandyopadhyay, B. Gain, A. Ekbal, A Neural Framework for English-Hindi Cross-Lingual Natural Language Inference, in: International Conference on Neural Information Processing, Springer, 2020, pp. 655–667.

[7] Y. Mehdad, M. Negri, M. Federico, Towards Cross-Lingual Textual Entailment, in: NAACL, pp. 321–324.

[8] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov, XNLI: Evaluating Cross-Lingual Sentence Representations, arXiv:1809.05053.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G.Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[10] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H. Huang, M. Zhou, InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training, 2021, arXiv:2007.07834.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805.

[12] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual Denoising Pre-Training for Neural Machine Translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742.

[13] Artetxe, Mikel, Ruder, Sebastian, Yogatama, Dani, On the cross-lingual transferability of monolingual representations, arXiv preprint arXiv:1910.11856.

[14] D. Khashabi, A. Cohan, S. Shakeri, P. Hosseini, P. Pezeshkpour, M. Alikhani, M. Aminnaseri, M. Bitaab, F. Brahman, S. Ghazarian,Parsinlu: A Suite of Language Understanding Challenges for Persian, arXiv preprint arXiv:2012.06154.

[15] H. Hu, H. Zhou, Z. Tian, Y. Zhang, Y. Ma, Y. Li, Y. Nie, K. Richardson, Investigating Transfer Learning in Multilingual Pre-trained Language Models through Chinese Natural Language Inference, arXiv preprint arXiv:2106.03983.

[16] K. T. K. Phan, D. N. Hao, D. Van Thin, N. L.-T. Nguyen, Exploring Zero-shot Cross-lingual Aspect-based Sentiment Analysis using Pre-trained Multilingual Language Models, in: 2021 International Conference on Multimedia Analysis and Pattern Recognition (MAPR), IEEE, 2021, pp. 1–6.

[17] N. T. Quyen, H. T. Anh, N. T. M. Huyen, N. Lien, VLSP 2021 - vnNLI Challenge: Vietnamese and English-Vietnamese Textual Entailment, VLSP 2021.

[18] S. Bowman, G. Angeli, C. Potts, C. D. Manning,A Large Annotated Corpus for Learning Natural Language Inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 632–642.

[19] A. Williams, N. Nangia, S. Bowman, A Broad-Coverage Challenge Corpus for Sentence Understanding Through Inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 ,2018, pp. 1112–1122.

[20] A. Conneau, G. Lample, Cross-Lingual Language

Model Pretraining, Advances in Neural Information Processing Systems, Vol. 32, 2019, pp. 7059–7069.

[21] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A Massively Multilingual Pre-TrainedText-To-Text Transformer, CoRR, arXiv:2010.11934.

[22] H. Hu, K. Richardson, L. Xu, L. Li, S. Ku¨bler, L. S. Moss, OCNLI: Original Chinese Natural Language Inference, arXiv:2010.05444.

[23] D. Pang, L. H. Lin, N. A. Smith, Improving Natural Language Inference with a Pretrained Parser, CoRR, arXiv:1909.08217.

[24] S. Wang, J. Jiang, Learning Natural Language Inference with LSTM, CoRR, abs/1512.08849. arXiv:1512.08849.

[25] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, arXiv:1705.02364.

[26] S. Chakravarthy, A. Umapathy, A. W. Black, Detecting Entailment in Code-Mixed Hindi-English Conversations, in: Proceedings of the Sixth Workshop on Noisy User-generatedText (W-NUT 2020), Association for Computational Linguistics, Online, 2020, pp. 165–170.

[27] A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, J. Goodfellow, Realistic Evaluation of Deep Semi-Supervised Learning Algorithms (2019). arXiv:1804.09170.

[28] D. V. Thin, L. S. Le, V. X. Hoang, N. L. T. Nguyen, Investigating Monolingual and Multilingual BERTModels for Vietnamese Aspect Category Detection, 2021, arXiv:2103.09519.

[29] H. Wu, S. Prasad, Semi-Supervised Deep Learning Using Pseudo Labels for Hyperspectral Image Classification, IEEE Transactions on Image Processing, Vol. 27, No. 3, 2018, pp. 1259–1270. doi:10.1109/TIP.2017.2772836.

[30] T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.

[31] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, arXiv preprint arXiv:1711.05101.