VNU Journal of Science:
Computer Science and Communication Engineering

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

Original Article

VLSP 2021 - vnNLI Challenge: Vietnamese and English-Vietnamese Textual Entailment

Ngo The Quyen[1,*], Hoang Tuan Anh[2], Nguyen Thi Minh Huyen[1], Nguyen Lien[2]

[1]*VNU University of Science, Hanoi, 334 Nguyen Trai, Hanoi, Vietnam*
[2]*FPT University, Hoa Lac, Hanoi, Vietnam*

Received 30 March 2022
Revised 9 April 2022; Accepted 5 May 2022

Abstract: This paper presents the first challenge on recognizing textual entailment (RTE), also known as natural language inference (NLI), held in a Vietnamese Language and Speech Processing workshop (VLSP 2021). The challenge aims to determine, for a given pair of sentences, whether the two sentences semantically agree, disagree, or are neutral/irrelevant to each other. The input sentences are in English or Vietnamese and may not be in the same language. This task is important in identifying, from different information sources, the evidence that supports or refutes a statement. The identification of such evidence is subsequently useful for many information tracking applications, such as opinion mining, brand and reputation management, and particularly fighting against fake news. Through this challenge, we would like to provide an opportunity for participants who are interested in the problem, to contribute their knowledge to improve the existing techniques and methods for the task, so as to enhance the effectiveness of those applications. In the paper, we introduce a collection of Vietnamese and English sentences in the domain of health that we built to serve as a benchmarking dataset for the task. We also describe the evaluation results of systems participating in the challenge.

Keywords: Natural Language Inference (NLI), Recognizing Textual Entailment (RTE).

1. Introduction

With the development of Internet and digital technologies, more and more content is created each day on social media and news or entertainment websites. The increasing amount of generated text data is both an opportunity and a challenge for researchers in the field of natural language processing (NLP). In this field, problems related to Natural Language Understanding (NLU) such as recognizing textual entailment (RTE) have been attracting a lot of attention from researchers worldwide. RTE, also known as natural language inference (NLI), is the task of determining whether a hypothesis is semantically true (entailment), false (contradiction), or undetermined (neutral), given a trusted premise.

* Corresponding author.
E-mail address: ngoquyenbg@hus.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.363

NLI, modeled as a sentence pair classification problem, is useful in many information tracking applications, such as opinion mining, machine translation, brand and reputation management, and particularly fighting against fake news. Fake news detection is a typical application that can use NLI solution in identifying, from large online information sources, the evidence that supports or refutes a statement [1]. For the problem of Aspect-Based Sentiment Analysis (ABSA), Sun et al [2] constructed auxiliary sentences from the aspects, converted this problem of sentence classification into a problem of sentence-pair classification, and achieved better results than systems previously published on the Semeval-2014 dataset. Another example is the application of NLI to the evaluation of machine translation systems, as it allows to determine the semantic correlation between two documents [3].

For English, the most studied language, there exist several datasets, both monolingual and multilingual, built for the NLI problem. The Stanford Natural Language Inference SNLI corpus for learning NLI [4], published in 2015, consists of 570K sentence pairs, written by humans based on captions from the Flickr30k corpus. The cross-lingual XNLI corpus dataset [5] (2018) is composed of 7,500 manually labeled English sentence pairs, translated into 14 languages (including Vietnamese), making a multilingual dataset of 112,500 annotated sentences pairs in total. Another famous source is GLUE, the General Language Understanding Evaluation benchmark [6], a collection of resources for training, evaluating, and analyzing NLU systems. GLUE contains 4 datasets related to NLI including MNLI (Multi-Genre NLI), QNLI (Question-answering NLI), WNLI (Winograd NLI), and RTE datasets made up of a series of RTE competitions. Recently, a new large scale benchmark dataset, Adversarial NLI (ANLI) [7], was collected via an iterative, adversarial human- and-model-in-the-loop procedure. Another dataset, called DocNLI [8], is composed of paragraph pairs instead of sentence pairs.

For Vietnamese, the NLI problem has not been well studied due to the lack of good datasets. To the best of our knowledge, XNLI is the only NLI dataset that involves Vietnamese sentences. Therefore, in VLSP 2021 we have decided to launch the first shared task for Vietnamese and English- Vietnamese Textual Entailment. The goal of this task is to produce a benchmark dataset for NLI with Vietnamese and English-Vietnamese sentence pairs, and to provide an opportunity for research groups to contribute their knowledge for developing high quality NLI systems and promoting NLP research for Vietnamese. The dataset includes 16,200 training sentence pairs and 4,177 test sentence pairs that are manually written by human subjects who are well educated. This NLI dataset is accessible for research purpose via the VLSP website.

The remainder of this report is organized as follows. Section 2 describes the shared task, the dataset construction and the evaluation measures. Section 3 introduces different approaches to the NLI problem. Section 4 summarizes and discusses about the participating systems and their results. Finally, we conclude the paper with some perspective.

Vietnamese is the official language of Vietnam with more than 76 million native speakers. It is the first language of the majority of the Vietnamese population. Several attempts have been conducted to build Vietnamese automatic speech recognition (ASR) system [1, 2, 3]. In 2013, the National Institute of Standards and Technology, USA released the Open Keyword Search Challenge (Open KWS) for Vietnamese speech. Many approaches have been proposed to improve performance for both keyword search and speech recognition [4-6].

Recently, The International Workshop on Vietnamese Language and Speech Processing (VLSP) has annually organized ASR challenge for Vietnamese. The VLSP Consortium1 regroups all academic and industrial research teams involved in Vietnamese language and speech processing. The first kick-off meeting of this community was in 2005 at the Institute of Information Technology, Vietnam Academy of

Science and Technology. The first ASR challenge was organized in VLSP 2018. In this challenge, no training dataset was released by the organizer. Participants used public or their own datasets to develop the models. Only 3 submissions were received. In VLSP 2019-ASR, a 500-hour-dataset was released by the organizer. However, participants could use any additional data to develop the models. In VLSP 2020-ASR, a 250-hour-dataset was released to participating teams to train the models. It was the first time the challenge was divided into two tasks. In Task1, participants had to only use training data provided by the organizer. In Task2, participants could use any resources to train their models. Finally, there were 10 submissions for Task1 and 4 submissions for Task2.

Note that, in all the previous ASR challenges, the data provided by the organizer were with manual transcription and not domain-specific. In the VLSP ASR 2021 challenge, we conducted more challenging and realistic tasks by focusing a specific domain i.e., online lectures. In addition, both labeled and unlabeled data were provided to participating teams. Specifically, the ASR challenge composed of two sub-tasks:

ASR-Task1 focuses on a full pipeline development of the ASR model from scratch. The organizer provided two training datasets. The first dataset is around 241.1 hours of transcribed data. Each participant had to label a part of the dataset before receiving the whole datasets. The second dataset is around 360.7 hours of untranscribed in-domain data. All participants were required to use only this provided data to develop models including acoustic and language models. Any use of another resource for model development was not acceptable.

ASR-Task2 focuses on spontaneous speech in different real scenarios e.g., meeting conversation, lecture speech. For this task, the organization did not provide training data, participants could use all available data sources to develop their models without any limitation. The ASR challenge attracted 47 registrations and 18 final result submissions. Many interesting approaches with remarkable results have been proposed by the participants. This paper presents the challenge description from data preparation to final result submission of participating teams. Moreover, different approaches for Vietnamese ASR will be described in details.

The rest of this paper is organized as follows. Section 2 provides information about participants and the processes in the challenge. Section 3 discusses the process of data preparation. Evaluation is described in Section 4. Finally, Section 5 concludes the paper.

2. Task Description

2.1. Task Definition

This challenge aims to determine, for a given pair of sentences, if the two sentences semantically agree, disagree, or are neutral/ irrelevant to each other. Here, the sentences are in English or Vietnamese and may not be in the same language. The input is a sentence pair sentence_1, sentence_2 pair, and the output is one of agree, disagree, neutral labels. The pair is agreed if sentence_2 can be inferred from sentence_1, disagree if two sentences have opposite meanings, and neutral if the two sentences are neither agree nor disagree though they may topically relevant to each other.

2.2. Data Collection

One of the important applications of NLI is fake news detection. As healthcare is a domain of great interest for most people, and a particular focus of attention of the whole society due to the COVID-19 pandemic, we choose to build a NLI benchmark dataset related to the medical field. We have collected data from a number of reputable news websites in Vietnamese and English, in the health category. Several criteria are applied to filter out the sentences in each article as “premise” sentences (sentence_1).

- The selected sentence is the first sentence of each paragraph;
- it does not contain question words;
- it should be sufficiently long, at least 10 tokens;

• it should not be more than 90% similar in words or syllables to an already selected sentence.

2.3. Building the NLI Dataset

Building NLI data implies more than simply labeling each pair of sentences: from the "premise" sentence, the annotator will have to write three "hypothesis" sentences corresponding to the three labels agree, disagree and neutral.

To make data construction more convenient and more efficient, we have built a tool for labeling NLI data. We have 33 annotators who are students from Vietnam National Universities in Hanoi and Ho Chi Minh city for writing hypothesis sentences, and 5 reviewers who are journalists and lecturers in Linguistics or in Journalism and Communications to evaluate sentences written by annotators. If a sentence is rated as unsatisfactory, the annotator will have to rewrite the sentence, until the reviewer accepts it as passing. During data construction, "premise" sentences that are judged to be inappropriate will also be discarded.

The NLI dataset is carefully constructed over all stages, with the purpose of building a good quality dataset for the competition as well as for the NLP research community.

The dataset for the competition includes 16,185 pairs of sentences for training and 4,177 pairs of sentences for testing.

2.4. Data Format

The data is provided in JSON format, each instance includes 6 main attributes as follows:

• id: unique id for the sentence pair

• lang_1: language of the first sentence, either 'vi' or 'en' for Vietnamese or English respectively

• lang_2: language of the second sentence, either 'vi' or 'en' for Vietnamese or English respectively

• sentence_1: the first sentence

• sentence_2: the second sentence

• label: a manually annotated label which marks the entailment relationship of the two sentences

– agree: If the two sentences semantically agree with each other.

– disagree: If the two sentences semantically disagree with each other.

– neutral: If the two sentences semantically neutral or irrelevant to each other.

Three examples of training data are given below:


```
{
    "id": "train_0",
    "lang_1": "en",
    "lang_2": "vi",
    "sentence_1": "Some Maine
    congressional leaders are pushing
    the SBA to amend a rule to help
    health care facilities.",
    "sentence_2": "Một số lãnh đạo
    quốc hội Maine đang tiến hành áp
    dụng các biện pháp phòng, chống
    COVID-19.",
    "label": "neutral"
}
```

```
{
    "id": "train_3",
    "lang_1": "en",
    "lang_2": "vi",
    "sentence_1": "Austin Regional
    Clinic is looking for 250
    volunteers to test a COVID-19
    vaccine by Pfizer Inc.",
    "sentence_2": "Các cuộc thử
    nghiệm vắc-xin ngừa virus corona
    chủng mới của Pfizer Inc trên
    người ở phòng khám khu vực
    Austin đang cần tuyển 250
    người tình nguyện.",
    "label": "agree"
}
```

```
{
    "id": "train_16",
    "lang_1": "vi",
    "lang_2": "vi",
    "sentence_1": "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.",
    "sentence_2": "Dù Tổng thống
    Trump đã dương tính với COVID-19
    nhưng vẫn chưa xuất hiện triệu chứng
    của bệnh.",
    "label": "disagree"
}
```


2.5. Evaluation Methods

The test data provided to the teams is a JSON file, which contains a list of instances. Each instance includes 3 attributes:
• id: unique id for the test sentence pair.
• sentence_1: the first sentence.
• sentence_2: the second sentence.


```
{
    "id": "test_0",
    "sentence_1": "Thống đốc Lamont
    đã thông báo vào chiều thứ Năm
    rằng bang Connecticut sẽ bắt đầu
    Giai đoạn 3 trong kế hoạch mở cửa
    trở lại.",
    "sentence_2": "Chiều thứ Năm,
    Thống đốc Lamont tuyên bố kích
    hoạt Giai đoạn 3 trong kế hoạch mở
    cửa trở lại ở bang Connecticut."
}
```


The result submission is a JSON file, which contains a list of instances. Each instance includes 2 attributes:
• id: unique id for the test sentence pair.
• label: a prediction label.
– agree
– disagree
– neutral


```
{
    "id": "test_0",
    "label": "agree"
}
```


The performances of NLI systems are evaluated by the F1-score for each type of label.

$$F_1 = \frac{2 \times P \times R}{(P + R)}$$

where P (Precision), and R (Recall) aredefined as follows:

$$P = \frac{SP_{true}}{SP_{sys}}$$

$$R = \frac{SP_{true}}{SP_{ref}}$$

where:

- $\boldsymbol{SP_{ref}}$: The number of **SP** (sentence pairs) in gold data;
- $\boldsymbol{SP_{sys}}$: The number of **SP** in recognizing system;
- $\boldsymbol{SP_{true}}$: The number of **SP** correctly predicted by the system.

The overall performance of the whole system is evaluated through accuracy, which is the proportion of correct predicted labels over the total number of predictions.

3. Approaches for NLI

Many methods have been proposed to solve the NLI problem. These approaches change over time, with the development of new methods and models being introduced, and can currently be divided into 3 main groups: symbolic (logic), statistical, and neural networks (deep learning) [9]. In the next part of this section, we will survey some approaches that have been applied to the NLI problem

3.1. Symbolic Approaches

Symbolic approaches use logical forms and processes to make inferences [9]. These approaches have primarily been applied in the early NLI challenges. For example, in [10], the authors present a system for textual inference that uses learning and a logical-formula semantic representation of the text. The system is built and evaluated based on the PASCAL RTE dataset [11]. Based on the dependent syntax, each sentence will be converted into a conjunction of logical terms, with the use of the cost function of an assumption to learn good assumption costs.

The approach in this study is a combination of statistical and classical logical reasoning. Another example is [12], where a system of logical inference which operates over natural language, called natural logic, is used for the NLI task. In the technical report of the fourth PASCAL recognizing textual entailment challenge [13] show that three teams (BOEING, Cambridge, and OAQA) approached the RTE problem using logical inference.

Methods based on logical inference do not require labeled datasets, but they do require expert knowledge to construct rules. Models built by these methods work relatively well on some data sets, but they lack generality and scalability.

3.2. Statistical Approaches

NLI can be considered as a classification problem, and many researchers approach this problem in the direction of building classifiers based on features. For example, in the report of the first TRE Challenge, [14] showed that The best system used a naıve Bayes classifier with features built from word co-occurrences. In the seventh RTE Challenge [15], a system with a similar approach also take the top spot. In [16], the authors proposed a method for RTE using lexical-level and sentence structure level features. The statistical measure of entailment between sentences is calculated based on acronyms (lexical level) extracted from the training data, and linguistic knowledge (sentence structure level).

Statistical-based methods have been applied to a wide variety of problems and give quite impressive results, but they often require a labeled dataset for training. Besides, these models often operate on a set of manual features, which has a great influence on the performance of the system.

3.3. Neural Network Approaches

The development of computing devices has allowed the creation of deeper and wider neural networks, and systems based upon those have achieved outstanding results in most NLP problems, including NLI. Along with the development of deep learning models, the advent of word embedding and document embedding methods as well as pretrained models also contributed significantly to the great advancement of NLP systems.

In [17], the authors introduced a new type of deep contextualized word representation named ELMo (Embeddings from Language Models). This word vectors model is pretrained on a large text corpus using a deep bidirectional language model (biLM), and significantly improves the state of the art across six challenging NLP problems, including question answering, textual entailment and sentiment analysis.

A significant breakthrough is introduced in [18], whose authors proposed a new language representation model named BERT (Bidirectional Encoder Representations from Transformers). This pretrained model can be fine-tuned to create state-of-the-art models for many NLP task such as question answering and language inference.

Neural network-based methods and especially deep learning have brought great strides in the performance of systems in various fields. End to end models or models based on pretrained datasets can create efficient systems without building feature sets. However, these approaches require huge computational resources as well as considerable training data sets.

New representation models give better representations of words, including in cases where a same word may have different representation vectors depending on the context in which it occurs. With the efficiency this model brings, more and more pretrained models are built for different languages, such as PhoBERT [19] for Vietnamese. This model is not only effective for monolingual data, but also works very well for multilingual data. Multilingual BERT (mBERT) and XLMR [20] are pretrained models that are used a lot in recent studies on NLP, including NLI problems. [21] present ALBERT (A Lite BERT), by using two parameter reduction techniques to lower memory consumption and increase the training speed of BERT; this model achieved the state of

the art results on many NLI datasets such as MNLI, QNLI, WNLI, RTE. NLI is a classification problem, so the combined use of machine learning techniques also contributes to increasing the performance of the model. Depending on the dataset and available computational resources, commonly used techniques such as learning rate adjustment (LRA), pseudo labels (PL), ensemble model, data augmentation, etc., may be used.

4. Submissions and Results

4.1. Submissions

49 teams pre-registered to participate in this NLI contest, among which 19 received data from the organizers. Finally, only 5 teams submitted results, and 4 of them submitted technical reports.

4.2. Technique and Resources

We named the four teams NLI1, NLI2, NLI3 and NLI4 respectively. The techniques and resources used by the teams are summarized in Table 1.

Team NLI1 builds models by fine-tuning pretrained models bert-base-multilingual- cased (mBERT) and xml-roberta-base (XLM-R). This team did not choose larger pretrained models due to limited computational resources, since they used thefree Google Colab environment.

Team NLI2 uses larger pretrained models like $XLM\text{-}R_{large}$ and $InfoXML_{large}$. In addition, this team also uses a combination of machine learning techniques such as Cross-validation (CV), Pseudo-labeling (PL) and Learning rate adjustment (LRA) to create a more robust model. Based on the experiments on the training dataset, the teamselected two models, InfoXLM and InfoXLM + LRA + PL to submit the results. Bothmodels are trained on Google Colab Pro environment.

Team NLI3 also uses two large pretrained models, $XLM\text{-}R_{large}$ and *RemBERT* [22]. This team created the biggest model of this year's competition. Instead of just using one last hidden layer, they concatenated the last four layers to form the representation vector for the classification model. The k- fold (with k = 5) cross-validation technique was used to generate 5 models based onXLM-R and 5 models based on RemBERT architecture. From these models, they create5 ensemble models. Each ensemble modelis made up of one XLM-R model and one RemBERT model, the ensemble models increasing efficiency by nearly 1% comparedto single models. In the prediction phase,the prediction probability is calculated as the average of the ouputs of the 5 ensemble models. Figures 1 and 2 show an overview ofthe model architecture of the team NLI3.

Table 1: Techniques used by the teams

Team	Pretrained models	Techniques
NLI1	mBERT, XLM-R	Just fine-tune on pretrained models
NLI2	InfoXLM	Data preprocessing, 10-fold cross validation Learning rate adjustment, Pseudo labels,
NLI3	RemBERT, XLM-R	5-fold cross validation, concatenate vector, Ensemble model
NLI4	XLM-R	Data preprocessing, 10-fold cross validation Data augmentation, Voting

Team NLI4 also performs fine tuning on the XLM-R model. This team also uses the k-fold cross-validation technique, and in addition, introduces data augmentation to enrich the data, increasing the size of the training data. The sentences in the training data are translated from Vietnamese to English and vice versa using Google Translation API. Experiments on the training dataset have shown that the use of data augmentation combined with word replacements has increased the performance of the model. This team used 10-fold cross-validation and the final result is obtained through voting between models.

The model training parameters of the teams are shown in Table 2.

4.2. Results

As described in the evaluation methods section, each team's results will include the F1 score for each label and the accuracy for the whole system.

Table 2: Model training parameters

Team	**Epoch**	**LR**	**Optimizer**	**Batch**
NLI1	30	2e-5	AdamW	8
NLI2	10	1e-5	AdamW	8
NLI3	5	2e-5	AdamW	12, 16
NLI4	3	1e-5	AdamW	16

Team NLI1 submitted 2 results, with the mBERT model getting an accuracy of 0.77 and the XLM-R model is 0.66.

Table 3 describes the results of team NLI1 with the mBERT model on the private test dataset. The NL1 model can be viewed as a baseline model.

Table 3: NLI1 mBERT base model result. (0-agree, 1- neutral, 2-disagree)

	precision	recall	F1-score	support
0	0.72	0.84	0.77	1394
1	0.83	0.82	0.82	1394
2	0.8	0.66	0.72	1389
acc	0.77			4177

Using a larger pretrained model, combined with machine learning techniques, team NLI2's model showed much better results. This team achieved an accuracy of 0.97 on the training dataset and 0.89 on the private test dataset. The results of team NLI2 are shown in Table 4.

Team NLI3 produced the largest model and also got the best result in this contest. Their model achieved an accuracy of 0.965 on the training dataset and 0.90 on the test dataset. The detailed results are presented in Table 5.

Team NLI4 achieved an accuracy of 0.88 on the test dataset, the results are presented in Table 6.

The results of the teams in this competition are summarized in Table 7.

Table 4: NLI2 result on private test data

	precision	recall	F_1-score	support
0	0.89	0.93	0.91	1394
1	0.92	0.88	0.90	1394
2	0.87	0.87	**0.87**	1389
acc	0.89			4177

Table 5: NLI3 result on private test data

	precision	recall	F_1-score	support
0	0.90	0.94	**0.92**	1394
1	0.91	0.91	**0.91**	1394
2	0.89	0.85	**0.87**	1389
acc	**0.90**			4177

Table 6: NLI4 results on private test data

	precision	recall	F1-score	support
0	0.89	0.92	0.90	1394
1	0.87	0.92	0.89	1394
2	0.90	0.81	0.85	1389
acc	0.88			4177

Table 7: Accuracy of systems

Team	**Accuracy**
NLI1	0.77
NLI2	0.89
NLI3	**0.90**
NLI4	0.88

5. Conclusion

In this paper, we have introduced the NLI problem and its applications, and the need for a NLI benchmark dataset for research in NLU for Vietnamese. We have built a NLI dataset of

more than 20,000 sentence pairs, both monolingual in Vietnamese and bilingual Vietnamese-English. The results obtained by the systems participing to the VLSP 2021 NLI challenge showed that use of pretrained models gives good results for the NLI problem. Larger pretrained models give better results, while combining models also increases the efficiency of the model, although not much. The best system presented for this year's competition achieved an overall accuracy of 90%, which is comparable to results obtained for English datasets.

With wide applicability of NLI in many NLP tasks, we plan to extends the datasets and continue this shared task in the next editions of VLSP workshop series.

Acknowledgements

This work is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14. We sincerely thank the data annotators and reviewers from the VNU University of Science, and VNUHCM University of Social Sciences and Humanities.

References

[1] K. C. Yang, T. Niven, H. Y. Kao, Fake News Detection as Natural Language Inference, in: 12th ACM International Conference on Web Search and Data Mining (WSDM-2019) (in Fake News Classification Challenge, WSDM Cup 2019), 2019.

[2] C. Sun, L. Huang, X. Qiu, Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence, in: NAACL, 2019.

[3] A. Poliak, Y. Belinkov, J. Glass, B. V. Durme,On the Evaluation of Semantic Phenomena in Neural Machine Translation Using Natural Language Inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2, 2018, pp. 513–523.

[4] S. R. 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. doi:10.18653/v1/D15-1075.

[5] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov, XNLI: Evaluating Cross-lingual Sentence Representations, in: EMNLP, 2018.

[6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A Multi-TaskBenchmark and Analysis Platform for Natural Language Understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks forNLP, Brussels, Belgium, 2018, pp. 353–355.

[7] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. eston, D. Kiela,Adversarial NLI: A New Benchmark for Natural LanguageUnderstanding, ArXiv abs/1910.14599.

[8] W. Yin, D. Radev, C. Xiong, DocNLI: A Large-scale Dataset for Document-level Natural Language Inference, 2021. arXiv:2106.09449.

[9] S. Storks, Q. Gao, J. Y. Chai, Commonsense Reasoning for Natural Language Understanding: A Survey of Benchmarks, Resources, and Approaches, ArXiv abs/1904.01172.

[10] R. Raina, A. Ng, C. D. Manning, Robust Textual Inference Via Learning and Abductive Reasoning, in: AAAI, 2005.

[11] I. Dagan, O. Glickman, B. Magnini, The PASCAL Recognising Textual Entailment Challenge, in: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché Buc (Eds.), Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, Springer, 2006, pp.177–190.

[12] B. MacCartney, C. D. Manning, Natural Logic for Textual Inference, in: ACL-PASCAL@ACL, 2007.

[13] D. Giampiccolo, H. T. Dang, B. Magnini, I. Dagan, E. Cabrio, W. B. Dolan, The Fourth PASCAL Recognizing Textual Entailment Challenge, in: TAC, 2008.

[14] I. Dagan, O. Glickman, B. Magnini, The PASCAL Recognising Textual Entailment Challenge, in: MLCW, 2005.

[15] L. Bentivogli, P. Clark, I. Dagan, D. Giampiccolo, The Seventh PASCAL Recognizing Textual Entailment Challenge, Theory and Applications of Categories.

[16] M. Tsuchida, K. Ishikawa, IKOMA at TAC2011: A Method for Recognizing Textual Entailment

using Lexical-level and Sentence Structure-level features, Theory and Applications of Categories, 2011.

[17] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, in: NAACL, 2018.

[18] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv: abs/1810.04805.

[19] D. Q. Nguyen, A. T. Nguyen, PhoBERT: Pre-trained language models for Vietnamese, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1037–1042.

[20] K. 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: ACL, 2020.

[21] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A LiteBERT for Self-supervised Learning of Language Representations, ArXiv: abs/1909.11942.

[22] H. W. Chung, T. Févry, H. Tsai, M. Johnson, S. Ruder, Rethinking embedding coupling in pre-trained language models, ArXiv: abs/2010.12821.