ViNLI: A Vietnamese Corpus for Studies on Open-Domain Natural Language Inference

Tin Van Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen
Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{tinhv,kietnv,ngannlt}@uit.edu.vn

Abstract

Over a decade, the research field of computational linguistics has witnessed the growth of corpora and models for natural language inference (NLI) for rich-resource languages such as English and Chinese. A large-scale and high-quality corpus is necessary for studies on NLI for Vietnamese, which can be considered a low-resource language. In this paper, we introduce ViNLI (**Vi**etnamese **N**atural **L**anguage **I**nference), an open-domain and high-quality corpus for evaluating Vietnamese NLI models, which is created and evaluated with a strict process of quality control. ViNLI comprises over 30,000 human-annotated premise-hypothesis sentence pairs extracted from more than 800 online news articles on 13 distinct topics. In this paper, we introduce the guidelines for corpus creation which take the specific characteristics of the Vietnamese language in expressing entailment and contradiction into account. To evaluate the challenging level of our corpus, we conduct experiments with state-of-the-art deep neural networks and pre-trained models on our dataset. The best system performance is still far from human performance (a 14.20% gap in accuracy). The ViNLI corpus is a challenging corpus to accelerate progress in Vietnamese computational linguistics. Our corpus is available publicly for research purposes[1].

1 Introduction

Although over 98 million people speak Vietnamese globally[2], Vietnamese is considered a low-resource language for natural language processing (NLP) research because of the lack of human-annotated corpora. To help accelerate NLP progress, Nguyen et al. (2020b, 2022) and Doan et al. (2021) collected a large number of human-annotated data to benchmark Vietnamese NLP tasks. We built the ViNLI corpus for evaluating natural language inference (NLI) models. NLI is an emerging and important task in natural language understanding which is to predict the semantic relation between two sentences. Several English corpora were released for the NLI task (Bowman et al., 2015; Williams et al., 2018). Recently, NLI has also witnessed corpus-creation efforts in other languages such as OCNLI (Hu et al., 2020), KorNLI (Ham et al., 2020), and IndoNLI (Mahendra et al., 2021). Quyen et al. (2022) proposed the Vietnamese-English NLI task[3] with three labels: agree, disagree, and neutral.

To contribute to the progress of NLP research for Vietnamese, we introduce a high-quality, open-domain corpus for Vietnamese NLI. Inspired by the success of NLI corpora (Bowman et al., 2015; Hu et al., 2020; Mahendra et al., 2021), we follow a strict annotation process and design the guidelines specific to Vietnamese characteristics to make the corpus realistic and high-quality. However, SNLI (Bowman et al., 2015) uses image captions as the main data resource, which are often short, simple texts and limited in linguistic phenomena. Therefore, we use a practical resource with various topics to capture diverse inferences of the Vietnamese NLI task. The premises in our corpus are sentences extracted from 800 news articles on 13 different topics.

In addition, most of the previous corpora require annotators to create only one hypothesis sentence for each premise, for example, SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), and XNLI (Conneau et al., 2018). However, in reality, human reasoning is very diverse in many semantic ways from a given sentence. Hence, we asked annotators to generate two hypothesis sentences for a premise sentence to capture many layers of semantic inference in our corpus. A Similar approach has been implemented in the OCNLI (Hu et al., 2020) and IndoNLI (Mahendra et al., 2021) corpora.

[1] UIT@NLP Group: https://nlp.uit.edu.vn/
[2] https://www.worldometers.info/world-population/vietnam-population/
[3] https://vlsp.org.vn/vlsp2021/eval/nli

Proceedings of the 29th International Conference on Computational Linguistics, pages 3858–3872
October 12–17, 2022.

ViNLI not only has three inference labels as in most previous corpora but also has one more label. The OTHER label is added to separate the inference types, which is different from the meaning of the NEUTRAL label because its purpose is to distinguish pairs of sentences that are unrelated in terms of semantic information, such as events, subjects, and objects. Table 1 shows several samples of Vietnamese NLI.

ViNLI is created and annotated by Vietnamese native speakers with solid linguistic backgrounds. Annotators are trained carefully to familiarize themselves with the corpus creation guidelines following a strict training process.

To evaluate the challenge of the corpus to models, we employ three deep neural networks, including CBoW (Mikolov et al., 2013), BiLSTM (Hochreiter and Schmidhuber, 1997), and ESIM (Chen et al., 2017). We also evaluate the SOTA pre-trained language models: BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and PhoBERT (Nguyen and Nguyen, 2020), which have achieved impressive performances on different NLP tasks.

Contributions of this study are as follows: (1) We introduce the ViNLI corpus, an open-domain, high-quality corpus consisting of over 30,376 human-annotated sentence pairs for evaluating the NLI task. (2) We conduct experiments on NLI models including neural network-based and pre-trained transformer-based models. (3) We analyze the corpus and the experimental results in different linguistic aspects to gain more insights into Vietnamese NLI.

2 Related Work

In this section, we review existing corpora and SOTA models for natural language inference.

2.1 Related Corpora

Early NLI corpora were created mainly for the task of Recognizing Textual Entailment (RTE) (Dagan et al., 2005; Toledo et al., 2012). These human-annotated corpora have contributed to evaluating statistical and logical NLI models. However, the main limitation of these corpora is the small size (less than a few thousand samples), which limits the ability to assess neural network-based models. For example, the SICK (Sentences Involving Compositional Knowledge) corpus (Marelli et al., 2014a) was used for the SemEval 2014 task with only 4,500 training samples. To overcome this limitation, the SICK corpus (Marelli et al., 2014b) was increased in scale to 10K samples.

Since 2015, many NLI benchmark corpora have been created to evaluate the effectiveness of machine learning models. In particular, Bowman et al. (2015) introduced SNLI, a first, large-scale, human-annotated corpus containing 570K English samples for evaluating NLI models. Then, a series of other large-scale NLI corpora appeared: STS-B (Cer et al., 2017) and QQP (Chen et al., 2018) for English with sizes of 8.5K and 404K samples, respectively. 2018 witnessed the release of two large-scale corpora, MultiNLI (Williams et al., 2018) comprising 433K samples, and XNLI (Conneau et al., 2018) with more than 112K samples. While the MultiNLI corpus was built for English, the XNLI was translated into 15 different languages. Besides, with the rapid growth of NLI research for English, NLP research communities witnessed the emergence of corpora for other languages, such as OCNLI (Hu et al., 2020) for Chinese, SICK-NL (Wijnholds and Moortgat, 2021) for Dutch, KorNLI (Ham et al., 2020) for Korean, IndoNLI (Mahendra et al., 2021) for Indonesian, NLI En-Hi (Khanuja et al., 2020) for Hindi-English, and FarsTail (Amirkhani et al., 2020) for Persian. Recently, the Adversarial NLI corpus was introduced by Nie et al. (2020), a data collection via human-and-model-in-the-loop training, which has brought new challenges to SOTA NLI models.

Quyen et al. (2022) introduced the bilingual (Vietnamese-English) NLI corpus[4] annotated with three labels: agree, disagree, and neutral, including approximately 16,200 sentence pairs in the medical domain. An open-domain, large-scale, high-quality corpus similar to SNLI or MultiNLI is necessary for Vietnamese NLI. Following corpus-development efforts, we create the ViNLI corpus, a high quality resource, for developing Vietnamese NLI models, which can improve other NLP tasks: machine translation, question answering, and text summarization.

2.2 Related Models

NLP has witnessed a rapid growth of large-scale corpora and deep learning models for NLI. Besides traditional machine learning models such as Skip-gram, CBOW (Mikolov et al., 2013), deep learning models such as RNN (Elman, 1990), Bi-RNN

[4]https://vlsp.org.vn/vlsp2021/eval/nli

Premise	Majority label All Labels Topic	Hypothesis
Hai cặp nam nữ bị cảnh sát bắt quả tang phê ma tuý nhảy múa trong tiếng nhạc công suất lớn ở căn hộ chung cư. *(Two male and female couples were caught by the police with narcotics and dancing to loud music in the apartment.)*	**Entailment** E E E E E Law	Có tổng cộng bốn người bị công an bắt giữ vì có hành vi sử dụng chất kích thích trái phép. *(A total of four people were arrested by the police for using illegal drugs.)*
Theo kế hoạch, Proace City Electric sẽ bán ra ở châu Âu từ cuối năm 2021. *(As planned, Proace City Electric will be sold in Europe from the end of 2021.)*	**Neutral** N N N N N Vehicles	Thị trường châu Âu vô cùng ưa chuộng dòng xe Proace City Electric. *(The European market extremely favors the Proace City Electric car.)*
Tương tự, đa số nhà đầu tư cá nhân cũng dự đoán giá vàng tăng. *(Similarly, the majority of individual investors are also predicting an increase in the price of gold.)*	**Contradiction** C C C C C Business	Giá vàng ngày càng giảm là điều đáng lo ngại được dự đoán bởi các nhà đầu tư cá nhân. *(The falling gold price is worrisome, which is predicted by individual investors.)*
Cổ phiếu UPS tăng hơn 10% khi lợi nhuận vượt dự báo của Wall Street. *(UPS shares rose more than 10% as earnings beat Wall Street forecasts.)*	**Other** O O O O O Business	NFT là một đơn vị dữ liệu trên sổ cái kỹ thuật số được gọi là blockchain. *(NFT is a data unit on a digital ledger called the blockchain.)*

Table 1: Several examples extracted from ViNLI with topic labels, gold inference labels and the four validation inference labels (E: Entailment, C: Contradiction, N: Neutral, and O: Other).

(Schuster and Paliwal, 1997), BiLSTM (Hochreiter and Schmidhuber, 1997), Dr-BiLSTM (Ghaeini et al., 2018), ESIM (Chen et al., 2017) have achieved positive results on NLI corpora. Recently, transformer-based models like BERT (Devlin et al., 2019) achieved significant progress in performance on various NLP tasks, including NLI on many well-known corpora, SNLI (Bowman et al., 2015) and MultiNLI (Liu et al., 2019a). Besides, the variants of BERT such as RoBERTa (Liu et al., 2019b), XLNet (Yang et al., 2019), and XLM-R (Conneau et al., 2020) also obtained outstanding results on the following corpora: MultiNLI (Williams et al., 2018), XNLI (Conneau et al., 2018), QQP (Chen et al., 2018), STS-B (Cer et al., 2017). However, the models have not been explored for Vietnamese NLI.

3 Corpus Creation

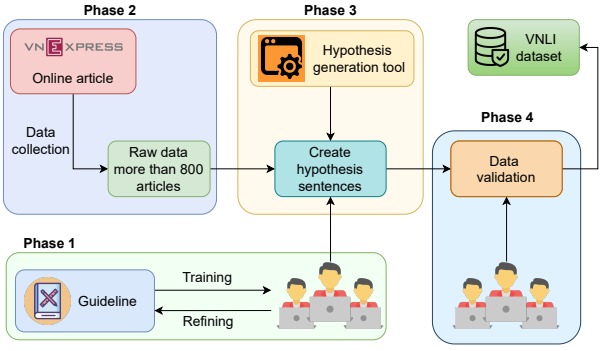


Figure 1: The process of corpus creation.

We build the ViNLI corpus following a strict process for quality control (see Figure 1). This process includes four primary phases: (3.1) annotator recruitment and training, (3.2) premise selection, (3.3) hypothesis generation, and (3.4) data validation. To obtain in-depth insights into the characteristics of the corpus, we analyze the corpus in terms of different linguistic aspects (see Section 3.5).

3.1 Annotator Recruitment and Training

Twenty-nine annotators with strong linguistic backgrounds contributed to the creation of premise-hypothesis pairs for the corpus. Figure 2 depicts our process of annotator training. The annotators must undergo this strict training phase before taking part in the official annotation process. First, the annotators are trained to familiarize themselves with the corpus-creation guidelines (see Section 3.3). Next, each annotator is required to create hypothesis sentences for the annotator-training set that contains 100 premise sentences. The labels (ENTAILMENT, CONTRADICTION, NEUTRAL, and OTHER) of their premise-hypothesis pairs on the set are masked. Two of the Twenty-eight other annotators are asked to give the labels for the same sentences. If the proportion of the labels agreed upon by the three annotators is over 0.95, the annotator is selected to participate in the official annotation process. Otherwise, that annotator needs to learn from their annotation mistakes and goes through the training phase again with another annotator-training set. Besides, we also discuss the annotation disagreements and identify complicated examples to refine the corpus-creation

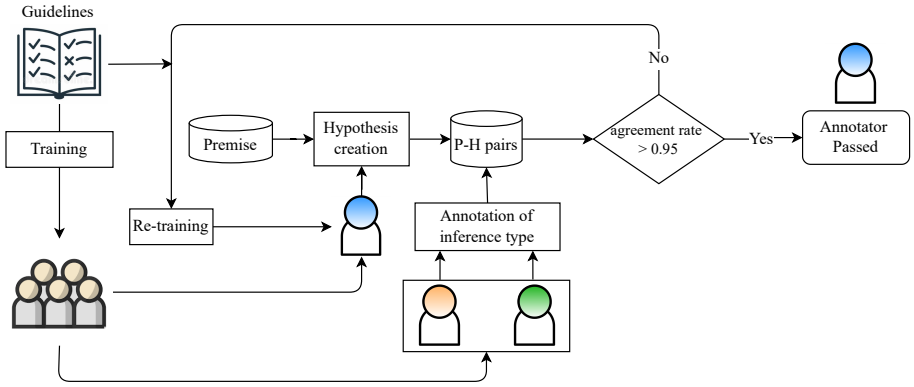


Figure 2: The strict process of training the annotators before creating our corpus.

guidelines.

3.2 Premise Selection

To capture the natural linguistic phenomena of Vietnamese NLI for news texts, the premises in our corpus ViNLI were extracted from 800 public articles on the reputable Vietnamese online newspaper VnExpress[5] with thirteen topics, shown in Table 9 in Appendix A. We selected articles with lengths of 3 to 5 paragraphs and chose the topic sentences of each paragraph as premise sentences. The reason we choose the topic sentence is because it clearly describes the main content of the whole paragraph and thus, is the most crucial sentence in a paragraph. As a result, each news article provides us with 3 to 5 premises.

3.3 Hypothesis Generation

We design four labels (ENTAILMENT, CONTRADICTION, NEUTRAL, and OTHER) instead of three (ENTAILMENT, CONTRADICTION, and NEUTRAL) as in SNLI, MultiNLI, or OCNLI datasets (Bowman et al., 2015; Williams et al., 2018; Hu et al., 2020). Because people also encounter pairs of sentences unrelated to each other in reality, and with only three labels is not possible to distinguish such cases. We ask annotators to create hypothesis sentences for the four following labels.

- **ENTAILMENT**: Create a hypothesis sentence that is definitely true to the content or situation of the premise sentence.
- **CONTRADICTION**: Create a hypothesis sentence that is definitely false with the content or situation of the premise sentence.
- **NEUTRAL**: Create a hypothesis sentence that might be true with the content or situation of the premise sentence.
- **OTHER**: Create a hypothesis sentence that is entirely unrelated to the content or situation of the premise sentence.

The OTHER hypothesis type is different from the NEUTRAL one. To create a NEUTRAL hypothesis, the annotators must rely on events, subjects, and objects in the premises. Along with this, annotators make the hypothesis that might be true with the content or situation of the premise. However, the OTHER hypothesis refers to an entirely different situation in the premise. Particularly, there is no connection of events, subjects, and objects between the premise and hypothesis.

Double Hypothesis: For each premise, annotators are required to create eight hypothesis sentences, two for each inference label (ENTAILMENT, CONTRADICTION, NEUTRAL, and OTHER). This strategy is similar to the multi-hypothesis strategy approached in the OCNLI and IndoNLI data creation protocol[6]. This data collection strategy requires annotators to mine the information in the premise sentence in many different semantic ways. In other words, the hypothesis sentence of each label has diverse content when the annotators are interested in many semantic aspects of the premises.

Besides, annotators were paid roughly 0.022 USD per premise-hypothesis pair. They must generate hypotheses based on the following guidelines:

[5] https://vnexpress.net/

[6] In OCNLI, three hypothesis sentences per label are created for each premise, resulting in a total of nine hypotheses. In IndoNLI, two hypothesis sentences per label are created for each premise, resulting in a total of six hypotheses

(1) Each premise has eight hypotheses (two hypotheses for each label). (2) Annotators are encouraged to write hypotheses in their own words without copying words or phrases from the premise. (3) Annotators may apply our general data-generation rules to create hypotheses. This rule set includes eight rules to generate CONTRADICTION sentences and eleven rules to create ENTAILMENT sentences, which are shown in Table 2 and Table 3 (see examples in Table 10 and Table 11 in Appendix D), respectively.

No.	Rule	Ratio
1	Use negative words (no, not, never, nothing, hardly, etc.)	22%
2	Replace words with antonyms	37%
3	Opposite of quantity	6%
4	Opposite of time	11%
5	Create a sentence that has the opposite meaning of a presupposition	11%
6	Wrong reasoning about an object (House, car, river, sea, person, etc.)	18%
7	Wrong reasoning about an event	27%
8	Others	4%

Table 2: Data-generation rules for creating CONTRADICTION hypothesis sentences.

No.	Rule	Ratio
1	Change active sentences into passive sentences and vice versa.	47%
2	Replace words with synonyms.	75%
3	Add or remove modifiers that do not radically alter the meaning of the sentence.	73%
4	Replace Named Entities with a word that stands for the class.	12%
5	Turn nouns into relative clauses	6%
6	Turn the object into relative clauses	7%
7	Turn adjectives into relative clauses	2%
8	Replace quantifiers with others that have a similar meaning.	13%
9	Create a presupposition sentence	8%
10	Create conditional sentences	2%
11	Others	2%

Table 3: Data-generation rules for creating ENTAILMENT hypothesis sentences.

3.4 Data Validation

To ensure the quality of annotating inference labels for premise-hypothesis pairs, we performed a round of data validation for the ViNLI corpus (see Table 9 in Appendix A). Each premise-hypothesis pair in the development and test dataset is annotated with inference labels by five different annotators. Annotators participating in the validation phase are those who joined in the hypothesis generation phase. The annotators who give inference labels must be different from the person who generate hypothesis in the hypothesis generation phase. Each premise-hypothesis pair is paid 0.013 USD.

We choose the final gold label for each premise-hypothesis pair by majority vote. Similar to the previous corpora (SNLI, MultiNLI, and OCNLI), if not at least three of five labels are the same for a pair, the gold labels are marked as '-'. And then, those pairs are either removed from the corpus or not used during model training and evaluation. The results of the validation phase are shown in Table 4. The statistics show 99.4% of sentence pairs receiving three or more identical labels, higher than the validation results of the well-known corpora such as SNLI, MultiNLI, and OCNLI, illustrating that the ViNLI has high quality and reliability.

3.5 Corpus Analysis

Before conducting corpus analysis, we divided ViNLI randomly into three sets: 80% for a training set (Train), 10% for a development set (Dev), and 10% for a test set (Test). Our corpus statistics are presented in Table 9 (in Appendix A).

The distribution of premise-hypothesis pairs and the average length of premise and hypothesis sentences (words) are illustrated in Table 9 (in Appendix A). We intentionally distribute the premise-hypothesis pairs of each topic evenly distributed on the Dev and Test sets. This division makes it possible to evaluate models fairly without bias toward any topic. In addition, the average length distribution of premise and hypothesis sentences in the Train, Dev, and Test sets is similar, with 24.5 words and 18.1 words for premise and hypothesis, respectively.

Length Distribution: The distribution of the premise and hypothesis sentences according to their length is shown in Figure 3 (in Appendix B). This Figure shows the same distribution as in Table 9. Most hypothesis sentences are shorter than premise sentences, which is similar to SNLI (Bowman et al., 2015). However, this indicates that the length of hypothesis sentences is shorter, but they still guarantee clear representations of their semantic reasoning from the sentence premises. The shortest and longest lengths to generate a hypothesis are 4 and 68 words, respectively. The number of hypothesis sentences with 10-23 word lengths occupies

Statistic	SNLI*	MultiNLI*	XNLI*	OCNLI†	IndoNLI§	ViNLI
Language	English	English	15 languages	Chinese	Indonesian	Vietnamese
Text genre	Image captions	Multi genre	Multi genre	Multi genre	Multi genre	Newswire
#pairs in total	570,152	432,702	7,500	56,525	17,736	30,376
#pairs relabeled	56,951	40,000	7,500	9,913	7,497	6,000
% relabeled	10.0%	9.2%	100%	17.5%	42.3%	19.8%
Pair w/unanimous gold label	58.3%	58.2%	nan	60.1%	nan	77.9%
4+ labels agree	nan	nan	nan	82.5	nan	91.5%
3+ labels agree	**98.0%**	**98.2%**	**93.0%**	**98.6%**	**98.6%**	**99.4%**
Individual label = gold label	89.0%	88.7%	nan	87.5%	90.0%	94.1%
Individual label = author's label	85.8%	85.2%	nan	80.8%	87.6%	91.1%
Gold label = author's label	91.2%	92.6%	nan	89.3%	92.3	96.4%
Gold label ≠ author's label	6.8%	5.6%	nan	9.3%	6.3	3.6%
No gold label (no 3 labels match)	2.0%	1.8%	nan	1.4%	1.4%	0.6%

Table 4: Agreement result of the validation phase in ViNLI compared with other corpora. *The numbers of SNLI, MultiNLI, XNLI corpora are extracted from the scientific papers (Bowman et al., 2015; Williams et al., 2018; Conneau et al., 2018). For XNLI, the number is calculated on a subset of Dev and Test in English. †For OCNLI, the number was calculated from Hu et al. (2020) by averaging 4 different protocols. §For IndoNLI, the agreement is calculated from Mahendra et al. (2021) by averaging 2 different protocols.

the largest proportion in our corpus.

Word Overlap: Taking the motivation from IndoNLI (Mahendra et al., 2021), we calculated the word overlap between the premise and hypothesis sentences in ViNLI. Higher word overlap rates can help predict the inference labels more correctly, which has been illustrated in the study of McCoy et al. (2019). Particularly, we use the Jaccard to calculate the unordered word overlap rate of premise-hypothesis pairs and the LCS index (the Longest Common Sub-sequence) to observe the level of word overlap in order. Before calculating the Jaccard and LCS, we used the VnCoreNLP toolkit (Vu et al., 2018) to perform word segmentation for Vietnamese, as shown in Table 5. With the Jaccard, the label ENTAILMENT has the highest rate of word overlap compared to the other labels, while this ratio is very low for the OTHER label. This is understandable because the hypothesis sentences of the OTHER label in ViNLI are created with content unrelated to its premise sentence. The word overlap rate in the order of premise and hypothesis in the ENTAILMENT label is also the highest, and the OTHER label is the lowest when compared with the CONTRADICTION and NEUTRAL labels. Compared with the IndoNLI corpus created by the lay annotators (Mahendra et al., 2021), the Jaccard and LCS are significantly lower, which can make ViNLI more interesting and challenging for evaluating NLI models.

New Word Rate: To evaluate word diversity in the hypothesis, we measure the new word rate, which is the proportion of hypothesis words not present in the premise. To detect Vietnamese words, we use the word segmentation tool VnCoreNLP (Vu et al., 2018). In Table 5, the new word rate in the ENTAILMENT hypotheses is the lowest at 46.59%. The word diversities of the CONTRADICTION (53.96%), NEUTRAL (61.79%), and OTHER (85.93%) labels are higher than that of the ENTAILMENT label. Compared with the IndoNLI corpus (Mahendra et al., 2021), the new word rate is remarkably higher, making ViNLI more diverse words to challenge NLI models.

In addition, we further analyze the tendency of using part-of-speech (POS) of the new words which annotators used to write hypotheses based on premises. Table 5 shows that annotators use nouns and verbs the most to create hypotheses. Before performing this statistic, we used PhoNLP (Nguyen and Nguyen, 2021) to identify the POS of words.

Data-Generation Rules Analysis: To understand the linguistic behaviors of annotators in creating ViNLI, we analyze the data-generation rules which annotators use to generate hypotheses. We randomly selected 100 ENTAILMENT premise-hypothesis pairs and 100 CONTRADICTION premise-hypothesis pairs in the corpus for analysis. For the CONTRADICTION label (see Table 2), the annotators use the "replace words with antonyms" rule with 37%, whereas the "opposite of quantity" rule is the lowest with 6%. For the ENTAILMENT label (see Table 3), "replace words with synonyms" and "add or remove modifiers that do not radically alter the meaning of the sentence" are the two most common rules used to generate

Label	Jaccard (%)	LCS	New word rate (%)	Part-Of-Speech (%)					
				Noun	Verb	Adjective	Preposition	Adjunct	Other
Entailment	29.88	52.90	46.59	31.45	24.97	6.67	8.39	8.71	19.81
Contradiction	23.30	48.90	53.96	30.79	23.53	7.40	7.61	11.27	19.40
Neutral	20.19	50.34	61.79	33.42	22.89	8.02	8.59	8.62	18.46
Other	6.18	45.53	85.93	42.16	21.91	7.73	8.02	4.87	15.31

Table 5: Word overlap between premise and hypothesis sentences.

hypotheses with 75% and 73%, respectively. While "create conditional sentences" rule is the least common to create hypotheses sentences, with only 2%. "Others" only accounts for a small part of our data.

During data generation, annotators may use one rule or more to generate a hypothesis. Our analysis on using rules for hypothesis creation (see Figure 5 in Appendix E) found that approximately two-thirds (66%) of the hypothesis sentences of the CONTRADICTION label are created by one rule, while data generation with two rules is about a third (32%). Surprisingly, very few annotators use more than two rules (only 2%) to generate hypotheses for the label CONTRADICTION. Unlike the CONTRADICTION label, the hypothesis sentences of the ENTAILMENT label (see Figure 6 in Appendix E) are usually created by combining two to three rules with 86%. Meanwhile, generating hypotheses using one rule or more than three rules for the ENTAILMENT label is only 14%.

4 Empirical Evaluation

4.1 Baseline Models and Settings

To evaluate the difficulty of ViNLI, we experiment with simple models (Random Guess and CBoW (Mikolov et al., 2013)) and more complex models (ESIM (Chen et al., 2017), BiLSTM (Hochreiter and Schmidhuber, 1997), PhoBERT (Nguyen and Nguyen, 2020), mBERT (Devlin et al., 2019), and XLM-R (Conneau et al., 2020)) as baseline models.

mBERT and XLM-R are multilingual language models pre-trained on multilingual documents, including Vietnamese. PhoBERT is a Pre-trained language model for Vietnamese that uses RoBERTa architecture with 135M parameters for the base version and 370M for the large version.

All baseline models are trained using Adam optimal function (Kingma and Ba, 2015) and on Tesla P100-PCIE-16GB of Google Colab[7]. In addition to identifying word boundaries, white space is also used to separate syllables that constitute words in Vietnamese texts. Models have different inputs, which can be word-based or syllable-based representations. Particularly, we implement the CBoW, ESIM, and BiLSTM models with the pre-trained embedding PhoW2V (Nguyen et al., 2020a) for Vietnamese. We experiment with two 300-dimensional versions of PhoW2V, including the syllable and word levels. When using PhoW2V with word-level, we use the VnCoreNLP toolkit (Vu et al., 2018) for word segmentation in Vietnamese.

The hyper-parameters of CBoW, ESIM, BiLSTM models are set up as follows: *learning_rate = 0.001, batch_size = 16, sequence_lenght = 80, epochs = 10*. To train transformer models like BERT, XLM-R, and PhoBERT, we used Hugging Face's Transformers library[8]. Besides, we set the hyper-parameters as follows *learning_rate = 1e-5, epochs = 10, batch_size = 16*.

We conduct experiments on two label sets: a three-label set (ENTAILMENT, CONTRADICTION, and NEUTRAL) and a four-label set (ENTAILMENT, CONTRADICTION, NEUTRAL, and OTHER).

4.2 Evaluation Metrics

Following SNLI (Bowman et al., 2015), we use accuracy as the primary evaluation metric. We also calculate F1-score (macro average) as the second evaluation metric to obtain more insights.

4.3 Human Performance

Following Hu et al. (2020), we hired five native Vietnamese speakers to annotate a subset of 300 samples (Test_{300}) extracted randomly from the Test set. These people did not know anything about the NLI task before, and we trained them on task definition and the meaning of inference labels to choose a label for each premise-hypothesis pair. The majority voting of five labels chooses the final label for each pair. Human performances are achieved with accuracy-based performances of

[7]https://colab.research.google.com/

[8]https://huggingface.co/docs/transformers/index

95.34% (for 03 labels: ENTAILMENT, CONTRADICTION, NEUTRAL) and 95.78% (for 04 labels: ENTAILMENT, CONTRADICTION, NEUTRAL, OTHER).

4.4 Experimental Results

Table 6 presents the performances of the baseline models on the Dev and Test sets. Overall, transformer-based models (mBERT, PhoBERT, and XLM-R) outperform others (Random Guess, CBOW, ESIM, and BiLSTM). XLM-R$_{Large}$ achieves the best results in both experiments with different label sets: three labels (ENTAILMENT, CONTRADICTION, NEUTRAL) and four labels (ENTAILMENT, CONTRADICTION, NEUTRAL, OTHER). On the three-label corpus, XLM-R$_{Large}$ achieved the highest accuracy-based performances on Dev and Test sets, with 83.02% and 81.36%, respectively. Besides, PhoBERT$_{Large}$ also achieves impressive results with 75.93% accuracy on the Test set.

On the corpus with four labels, the performances of the models tend to be quite similar to the three labels experiments. The performance of the XLM-R$_{Large}$ model also has the highest accuracy, with 86.77% on the Dev set and 85.99% on the Test set. The best performance of the syllable-level model (XLM-R) is significantly higher than the best performance of the word-level model (PhoBERT), similar to the Vietnamese MRC shared task (Nguyen et al., 2022). The accuracy and F1 achieve roughly the same results due to the experiments on the balanced corpus. Furthermore, most of the model performances on the four-label experiments are higher than those on the three-label experiments because the OTHER label is easier to recognize than other labels (see Table 7).

XLM-R$_{Large}$ also achieves the best accuracy on the Test$_{300}$ set. However, when compared with human performance, this model is still significantly lower, by 14.20% on Test$_{300}$ on three labels and 6.93% on Test$_{300}$ on four labels.

5 Result Analysis

To gain more insights, we analyze the two best-performance models, including XLM-R$_{Large}$ and PhoBERT$_{Large}$ on different linguistic aspects. In this section, XLM-R and PhoBERT stand for XLM-R$_{Large}$ and PhoBERT$_{Large}$, respectively.

How do inference labels affect the performance?

Table 7 shows the analysis of accuracy on each label in the Dev set on three labels and the Dev set on four labels of the two best-performance models (XLM-R and PhoBERT). Both XLM-R and PhoBERT perform very well on the OTHER label with more than 97% accuracy. When adding the OTHER label, the accuracy-based performances of XLM-R on CONTRADICTION and NEUTRAL are improved; however, the accuracy of ENTAILMENT is decreased from 89.31% to 86.33%.

Labels	Three-label Dev		Four-label Dev	
	PhoBERT	XLM-R	PhoBERT	XLM-R
Entailment	77.94	89.31	76.45	86.33
Contradiction	76.57	80.76	71.46	82.98
Neutral	77.53	79.12	77.66	80.45
Other	-	-	97.35	97.34

Table 7: Model performance per label in ViNLI.

How do new words affect the performance? We aim to analyze the effect of the new words in hypothesis sentences on the model accuracy. Figure 4 (in Appendix C) shows the accuracy of PhoBERT and XLM-R models according to the new word rate in the three-labels and four-labels dev set. Figure 4a shows that the accuracy of the PhoBERT model significantly decreases from around 84% to about 67% as the new word rate in hypothesis sentences increases from less than 20% to more than 80%. In contrast, the accuracy of the XLM-R model is relatively stable.

With the results of the four-label Dev set in Figure 4b, the accuracy of the models is quite similar to the trend of the three-label dev set when the new word rate is less than or equal to 60%. However, the performance of the models with the new word rate of more than 60% on the four-label Dev set is higher than that of the three-label Dev set since all pairs of OTHER labels have the highest new word rate (see Table 5). Moreover, the model performances on OTHER achieve the most (see Table 7) compared to other labels.

How do data-generation rules affect the performance? Table 8 analyzes the influence of using data-generation rules that the annotators generate the hypotheses on the model performance. We analyze our experimental results of PhoBERT and XLM-R on data-generation rules (as described in Subsection 3.5). Our experiments show that the ENTAILMENT hypotheses with more than two rules cause the performance of the models to be lower when these hypotheses are generated with only one rule. With the CONTRADICTION hy-

Model		Three Labels						Four Labels					
		Dev		Test		Test_{300}		Dev		Test		Test_{300}	
		Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
Random Guess		33.36	32.49	32.51	33.01	34.28	34.27	25.47	24.27	24.51	24.85	25.19	25.17
Syllable	CBoW	45.54	45.13	44.96	44.62	43.86	43.43	46.29	45.73	45.97	45.46	43.58	43.88
	ESIM	48.55	48.24	47.43	46.76	46.92	46.41	48.58	48.45	47.44	47.17	43.58	42.99
	BiLSTM	48.07	48.10	46.42	46.35	46.92	46.84	48.89	48.36	48.55	48.06	48.31	48.07
	mBERT	67.41	67.46	64.84	64.83	64.91	64.85	73.91	73.83	73.45	73.62	75.34	75.78
	XLM-R_{Base}	72.02	71.99	71.59	71.51	71.93	71.55	76.97	76.93	76.83	77.01	77.71	78.31
	XLM-R_{Large}	**83.02**	**82.98**	**81.36**	**81.31**	**81.14**	**81.12**	**86.77**	**86.76**	**85.99**	**86.10**	**88.85**	**89.13**
Word	CBoW	49.05	48.64	45.80	45.41	42.11	41.37	54.63	54.48	53.12	52.87	51.01	50.60
	ESIM	49.84	49.75	48.18	48.12	40.79	40.25	48.68	47.99	48.37	47.83	46.95	46.14
	BiLSTM	48.91	48.71	46.77	46.59	43.42	42.52	50.48	49.89	49.78	49.22	49.32	48.31
	PhoBERT$_{Base}$	75.07	75.08	72.87	72.79	71.05	70.31	79.79	79.75	78.00	78.05	77.70	77.98
	PhoBERT$_{Large}$	**77.33**	**77.34**	**75.93**	**75.87**	**77.19**	**77.19**	**80.72**	**80.72**	**80.67**	**80.69**	**80.74**	**81.11**
Human performance		-	-	-	-	**95.34**	**95.33**	-	-	-	-	**95.78**	**95.79**

Table 6: Human and machine performances on the Dev and Test sets of our corpus ViNLI.

potheses, the models have more difficulty than those generated by the annotator with only one rule, whereas PhoBERT and XLM-R achieve higher accuracy if the hypotheses are generated from two or more rules.

Label	**#Rule**	**PhoBERT**	**XLM-R**
Entailment	1-2	81.82	89.09
	more than 2	77.78	84.44
Contradiction	1	71.21	81.81
	more than 1	76.47	91.18

Table 8: Effects of data-generation rules on the models.

How do multiple topics affect the performance? To observe the impact of multiple topics (open-domain) in ViNLI, we calculate the accuracy of the two highest-performance models (PhoBERT and XLM-R models) in terms of 13 different topics. Table 12 (in Appendix F) shows that XLM-R consistently outperformed PhoBERT on all topics. The models achieve different results on various topics. While the models reach the best performances in Tourism and Entertainment, Business and World are the most challenging for the models.

6 Conclusion and Future Work

In this paper, we introduced ViNLI, an open-domain, high-quality corpus for evaluating Vietnamese NLI models. By a strict annotation scheme with high annotator agreements, 30,376 premise-hypothesis pairs of the corpus were annotated by humans with solid linguistic backgrounds, which is the largest Vietnamese NLI corpus to date. The performances of the two powerful models (XLM-R and PhoBERT) illustrate that our corpus is challenging for the pre-trained language models in Vietnamese, the best of which underperforms humans by over 14% (on three labels). ViNLI is available freely for research purposes in developing Vietnamese NLU models.

Taking advantages of state-of-the-art models on large-scale, high-quality NLI corpora (SNLI, MultiNLI, XNLI, OCNLI, KorNLI, and IndoNLI), we hope that our corpus will accelerate progress on Vietnamese NLI and other NLP tasks. Based on the findings of our work, we continue to enhance the quality and quantity of the corpus with different data sources and expand our corpus with adversarial samples (Kang et al., 2018; Nie et al., 2020). Moreover, future studies will concentrate on exploiting Vietnamese models using BERTology (Rogers et al., 2020) (e.g., SBERT (Reimers and Gurevych, 2019)) and improving NLP applications such as Vietnamese machine reading comprehension models and retriever-reader question answering systems.

Acknowledgements

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS2022-26-01. Tin Van Huynh was funded by Vingroup JSC and supported by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Institute of Big Data, code VINIF.2021.ThS.49. We would like to thank anonymous reviewers from ACL Rolling Review and COLING 2022, which their comments help us improve the quality of our paper. In addition, we would like to thank our annotators for their cooperation.

References

Hossein Amirkhani, Mohammad Azari Jafari, Azadeh Amirak, Zohreh Pourjafari, Soroush Faridan Jahromi, and Zeinab Kouhkan. 2020. Farstail: A persian natural language inference dataset. *arXiv preprint arXiv:2009.08820*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs. *URL https://www.kaggle. com/c/quora-question-pairs*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina 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 Natural Language Processing*, pages 2475–2485.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Long Doan, Nguyen Luong Tran, Thai Hoang, Dat Quoc Nguyen, et al. 2021. Phomt: A high-quality and large-scale benchmark dataset for vietnamese-english machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z Fern, and Oladimeji Farri. 2018. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. *arXiv preprint arXiv:1802.05577*.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for korean natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 422–430.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3512–3526.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 2018. Adventure: Adversarial training for textual entailment with knowledge-guided examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2418–2428.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. in yoshua bengio and yann lecun editors. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. Indonli: A natural language inference dataset for indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527.

M Marelli, Luisa Bentivogli, M Baroni, R Bernardi, Stefano Menini, and R Zamparelli. 2014a. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014b. A sick cure for the evaluation of compositional distributional semantic models. In *Lrec*, pages 216–223. Reykjavik.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020a. A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020b. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linh The Nguyen and Dat Quoc Nguyen. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–7.

Van Kiet Nguyen, Son Quoc Tran, Luan Thanh Nguyen, Tin Van Huynh, Son T Luu, and Ngan Luu-Thuy Nguyen. 2022. Vlsp 2021 shared task: Vietnamese machine reading comprehension. *arXiv preprint arXiv:2203.11400*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Ngo The Quyen, Hoang Tuan Anh, Nguyen Thi Minh Huyen, and Nguyen Lien. 2022. VLSP 2021 - vnNLI Challenge: Vietnamese and English-Vietnamese Textual Entailment. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Assaf Toledo, Sophia Katrenko, Stavroula Alexandropoulou, Heidi Klockmann, Asher Stern, Ido Dagan, and Yoad Winter. 2012. Semantic annotation for textual entailment recognition. In *Mexican International Conference on Artificial Intelligence*, pages 12–25. Springer.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Gijs Wijnholds and Michael Moortgat. 2021. Sick-nl: A dataset for dutch natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1474–1479.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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, Volume 1 (Long Papers)*, pages 1112–1122.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A Topic Statistics

Table 9 presents statistics in terms of 13 topics, mean premise length (words)[9] and mean hypothesis length (words)[10].

Topic/Label	Train	Dev	Test	Total
Technology	1,912	232	232	2,376
Tourism	1,896	232	228	2,356
Education	1,936	232	231	2,399
Entertainment	1,640	231	231	2,102
Science	1,792	232	228	2,252
Business	1,616	231	228	2,075
Law	1,680	231	231	2,142
Health	2,048	231	232	2,511
World	2,040	232	229	2,501
Sports	2,088	230	231	2,549
News	1,576	231	231	2,038
Vehicles	2,288	232	228	2,748
Life	1,864	232	231	2,327
Entailment	6,094	739	750	7,583
Contradiction	6,094	764	737	7,595
Neural	6,094	752	777	7,623
Other	6,094	754	727	7,575
Total (pairs)	24,376	3,009	2,991	30,376
MPL (words)	24.5	24.6	24.3	24.5
MHL (words)	18.3	17.9	18.1	18.1

Table 9: ViNLI statistics in terms of different topics, Mean Premise Length (MPL) and Mean Hypothesis Length (MHL).

B Length Distribution

The distribution of the premise and hypothesis sentences according to their length is shown in Figure 3. The length of premise and hypothesis sentences is counted by the number of words (A Vietnamese word consists of one or more syllables). We use the VnCoreNLP toolkit (Vu et al., 2018) for Vietnamese word segmentation.

[9]Mean premise length is the mean average of word-based lengths of premise sentences in Train/Dev/Test sets or ViNLI (total).

[10]Mean hypothesis length is the mean average of word-based lengths of hypothesis sentences in Train/Dev/Test sets or ViNLI (total).

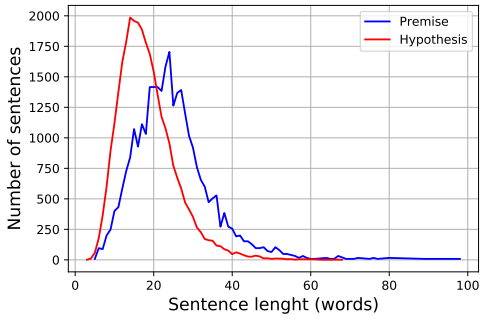
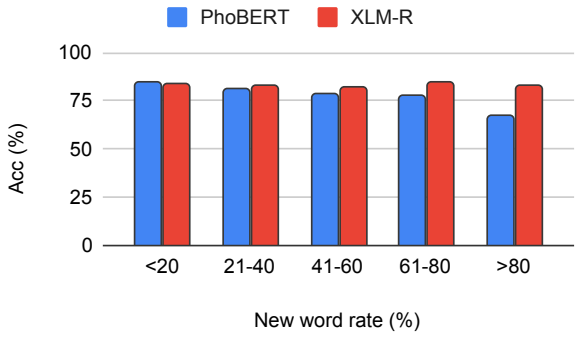


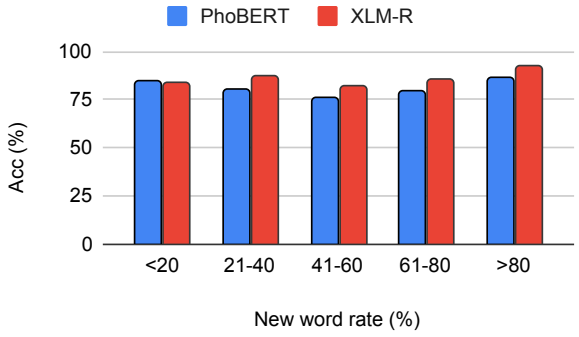
Figure 3: The distribution of sentence length.

C Effect of New Words

To observe the influence of the new word rate on the performance of models, we analyze the accuracy of PhoBERT$_{Large}$ and XLM-R$_{Large}$ models according to the new word rate. The analysis of the three-label and four-label Dev sets is shown in Figure 4.



(a) Dev set with three labels.



(b) Dev set with four labels.

Figure 4: Model accuracy on the Dev set according to new word rate.

D Data-Generation Rules

To understand the linguistic behaviors of annotators in creating ViNLI, we analyze data-generation rules which annotators use to generate hypothe-

ses. We randomly selected 100 ENTAILMENT premise-hypothesis pairs and 100 CONTRADICTION premise-hypothesis pairs in the corpus for analysis. For the CONTRADICTION label (see Table 2), the annotators use the "replace words with antonyms" rule with 37%, whereas the "opposite of quantity" rule is the lowest with 6%. For the ENTAILMENT label (see Table 3), "replace words with synonyms" and "add or remove modifiers that do not radically alter the meaning of the sentence" are the two most common rules used to generate hypotheses with 75% and 73%, respectively. While "create conditional sentences" rule is the least common to create hypotheses, with only 2%. "Others" only accounts for a small part of our data.

E Rules Combination for Creating Hypotheses

To create more diverse and challenging data, annotators may use one rule or more to generate a hypothesis. Figure 5 and Figure 6 show the proportion of using data-generation rules to create sentences for contradiction and entailment, respectively. Whereas most contradiction hypothesis sentences use one rule, entailment hypothesis sentences are created mainly based on two and three rules. Table 10 and Table 11 present several samples of contradiction and entailment rules for creating premise (P) - hypothesis (H) pairs, respectively.

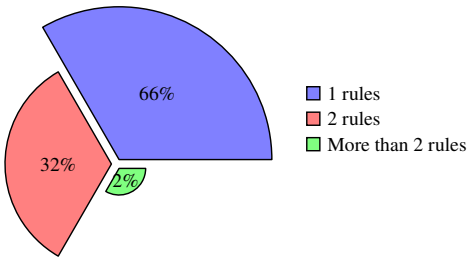


Figure 5: The ratio of combining different rules to create contradiction sentences in ViNLI.

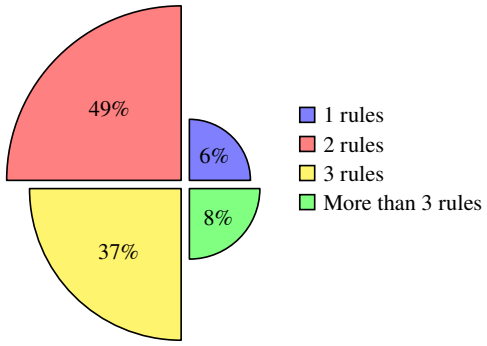


Figure 6: The ratio of combining different rules to create entailment sentences in ViNLI.

F Effects of Multiple Topics

To observe the impact of multiple topics (open-domain) in the ViNLI corpus, we calculate the accuracy of the two highest-performance models (PhoBERT$_{Large}$ and XLM-R$_{Large}$ models) on 13 different topics of ViNLI. The results are shown in Table 12.

Topic	Three-label Dev		Four-label Dev	
	PhoBERT	XLM-R	PhoBERT	XLM-R
Technology	76.88	83.24	77.59	86.64
Tourism	**84.88**	**89.53**	81.47	**91.81**
Education	79.43	81.14	82.76	83.62
Entertainment	81.14	88.57	**83.12**	90.91
Science	73.10	79.53	78.88	85.34
Business	**72.09**	80.23	**75.32**	**80.95**
Law	77.71	81.14	82.25	85.28
Health	80.92	83.82	82.68	88.74
World	72.99	**79.31**	79.31	87.93
Sports	75.86	85.63	83.04	83.91
News	76.30	81.50	82.68	88.31
Vehicles	78.16	83.33	81.90	87.93
Life	75.86	82.18	78.45	86.64

Table 12: Analyzing the model performances on different topics.

Rule	Example	Per.
Use negative words (no, not, never, nothing, hardly, etc.)	**P**: Cơ quan chức năng đã lập biên bản vụ việc. *(Authorities recorded the minutes of the incident.)* **H**: Cơ quan chức năng **không** tiến hành xử lí vụ việc. *(Authorities did **not** process the case.)*	22%
Replace words with antonyms	**P**: AAPP cho biết thêm chính quyền quân sự Myanmar đang **giam** 4.120 người, trong đó có 20 người bị kết án tử hình. *(The AAPP added that Myanmar's military junta is **holding** 4,120 people, of which 20 are sentenced to death.)* **H**: Đã có 4.120 người được chính quyền quân sự quân sự Myanmar **trả tự do**. *(There have been 4,120 people **released** by the military junta of Myanmar.)*	37%
Opposite of quantity	**P**: Suốt cuộc diễu hành kéo dài khoảng **một tiếng**, Tổng thống Bolsonaro và đa số người ủng hộ ông đều không đeo khẩu trang. *(During the parade, which lasted about **an hour**, President Bolsonaro and most of his supporters were not wearing masks.)* **H**: Cuộc diễu hành kéo dài khoảng **10 tiếng**, người tham gia và cả Tổng thống Bolsonaro đều không đeo khẩu trang. *(The parade lasted about **10 hours**; participants and President Bolsonaro were not wearing masks.)*	6%
Opposite of time	**P**: Miss Universe 2020 kéo dài khoảng 12 ngày, chung kết diễn ra tối **16/5** tại Hollywood, bang Florida. *(Miss Universe 2020 lasts about 12 days, and the final will take place on the evening of **May 16** in Hollywood, Florida.)* **H**: Miss Universe 2020 sẽ được tổ chức trong khoảng thời gian **từ 20-25/5** tại Mỹ. *(Miss Universe 2020 will be held **between May 20 and 25** in the US.)*	11%
Create a sentence that has the opposite meaning of a presupposition	**P**: **Giám đốc điều hành Apple, Lisa Jackson**, cho biết khó khăn của việc sử dụng năng lượng sạch. *(**Apple CEO Lisa Jackson** said the difficulty of using clean energy.)* **H**: Apple **chưa thể bổ nhiệm ai cho chức vụ giám đốc điều hành**. *(Apple **has not been able to appoint anyone for the position of CEO**.)*	11%
Wrong reasoning about an object (House, car, river, sea, person, etc.)	**P**: Trong thông báo hôm 24/5, **Honda Việt Nam** công bố chiến dịch thu hồi mẫu xe ga nhập khẩu. *(In an announcement on May 24, **Honda Vietnam** announced a campaign to recall imported scooter models.)* **H**: **Honda Hàn Quốc** thông báo triệu hồi các mẫu xe ga nhập khẩu. *(**Honda Korea** announced the recall of imported scooter models.)*	18%
Wrong reasoning about an event	**P**: Trong lần gặp lại này, **Zverev vượt trội đối thủ** ở giao bóng. *(In this meeting, **Zverev outperformed his opponent** in serving.)* **H**: **Đối thủ có kỹ năng giao bóng vượt xa** Zverev. *(**The opponent's serving skill far exceeded** Zverev.)*	27%
Others	**P**: Phía Tập đoàn Trung Nam cho biết, với 64,9% cổ phần còn lại họ vẫn giữ vai trò quyết định trong điều hành, định hướng phát triển của dự án điện gió Trung Nam. *(Trung Nam Group said that with the remaining 64.9% stake, they still play a decisive role in the management and development orientation of the Trung Nam wind power project.)* **H**: **Do** nắm giữ ít cổ phần **nên** Trung Nam Group mất quyền quyết định đối với các dự án quan trọng. *(**Because of** holding fewer shares, Trung Nam Group lost decision-making power on important projects.)* **Explanation**: Causal, Although clauses, etc., clauses can be used to create the contradiction hypothesis sentence with the premise sentence. This is a case of Others.	4%

Table 10: Examples of contradiction rules for creating premise (**P**) - hypothesis (**H**) pairs. Simply, we only mention one rule to be applied in each example.

Rule	Example	Per.
Change active sentences into passive sentences and vice versa.	**P**: Giá các mặt hàng dầu **đều tăng**. (*Prices of oil commodities **have increased**.*) **H**: Giá xăng dầu **được** tất cả các cửa hàng xăng dầu trên toàn quốc **điều chỉnh tăng lên**. (*Oil prices **are adjusted to increase by** all petrol stations nationwide.*)	47%
Replace words with synonyms.	**P**: Nadal tốn 130 phút để **vượt qua** Sinner 7-5, 6-4. (*Nadal took 130 minutes to **beat** Sinner 7-5, 6-4.*) **H**: Sau hơn hai tiếng, Nadal **chiến thắng** trước Sinner với tỷ số 2-0. (*After more than two hours, Nadal **won** against Sinner with a score of 2-0.*)	75%
Add or remove modifiers that do not radically alter the meaning of the sentence.	**P**: Châu Nhuận Phát **sinh ngày 18/5/1955** trong gia đình nghèo. (*Chau Nhuan Phat **was born on May 18, 1955** in a poor family.*) **H**: Châu Nhuận Phát **là con của một gia đình** có hoàn cảnh khó khăn. (*Chau Nhuan Phat **is the son of a family** with difficult circumstances.*)	73%
Replace Named Entities with a word that stands for the class.	**P**: Hacker rao bán dữ liệu của hơn 533 triệu **tài khoản Facebook**, bao gồm số điện thoại và một số thông tin cá nhân. (*Hacker sells data of more than 533 million **Facebook accounts**, including phone numbers and some personal information.*) **H**: Thông tin cá nhân của hơn 533 triệu **tài khoản mạng xã hội** đã bị Hacker rao bán. (*Personal information of more than 533 million **social network accounts** is sold by Hackers.*)	12%
Turn nouns into relative clauses	**P**: **Tháng 8-9** là thời điểm ốc béo nhất. (***August-September** is the fattest time of snails.*) **H**: **Tháng 8-9 là mùa mà** những người đầu bếp sẽ dễ dàng lựa chọn những con ốc béo và thơm nhất. (***August-September is the season when** chefs will easily choose the fattest and most fragrant snails.*)	6%
Turn the object into relative clauses	**P**: Wernery cho biết lạc đà được tiêm **xác của nCoV** để sản sinh kháng thể. (*Wernery said camels are injected with **the carcass of nCoV** to produce antibodies.*) **H**: Lạc đà được tiêm **xác của nCoV, là dung dịch** có khả năng tạo ra khoáng thể chống lại virus. (*Camels are injected with **the carcass of nCoV, which is a solution** capable of creating antibodies against the virus.*)	7%
Turn adjectives into relative clauses	**P**: Quần đảo Lofoten của Na Uy là một trong những **địa điểm đẹp nhất trên trái đất**. (*Norway's Lofoten Islands are some of **the most beautiful places on earth**.*) **H**: Quần đảo Lofoten của Na Uy là **địa điểm du lịch, nơi được mệnh danh là đẹp nhất trên trái đất**. (*Norway's Lofoten Islands are **tourist destinations that have been dubbed the most beautiful place on earth**.*)	2%
Replace quantifiers with others that have a similar meaning.	**P**: Công an xác minh, giờ ra chơi sáng 13/5, **một nam sinh và một nữ sinh** lớp 9B trong lúc đùa nghịch đã cắn tay nhau. (*Police verified that at recess on the morning of May 13, **a male student and a female student** in class 9B bit each other's hands while frolicking.*) **H**: **Hai học sinh** của lớp 9B đã cắn nhau vào giờ ra chơi. (***Two students** from class 9B were biting each other during break time.*)	13%
Create a presupposition sentence	**P**: Cũng theo Goal, **Marcelo không phải là cầu thủ duy nhất của Real** bất bình với Zidane. (*According to Goal, **Marcelo is not the only player of Real** to be angry with Zidane.*) **H**: **Marcelo đá cho đội tuyển Real Madrid**. (***Marcelo plays for the Real Madrid team**.*)	8%
Create conditional sentences	**P**: Do ảnh hưởng của Covid-19, doanh nghiệp không xuất khẩu được nên khoảng 50.000 tấn hành tím tới kỳ thu hoạch của nông dân xã Vĩnh Châu không có nơi tiêu thụ. (*Due to the impact of Covid-19, businesses could not export, so about 50,000 tons of purple onions until the harvest period of Vinh Chau commune farmers have no place to consume.*) **H**: **Nếu** không bị ảnh hưởng bởi Covid-19, doanh nghiệp **sẽ** xuất khẩu được khoảng 50.000 tấn hành tím. (***If** not affected by Covid-19, the enterprise **will** be able to export about 50,000 tons of purple onions.*)	2%
Others	**P**: Nạn nhân không bị nguy hiểm đến tính mạng nhưng chưa thể làm việc với cơ quan điều tra. (*The victim's life is not in danger, but the victim has not been able to work with the investigative agency.*) **H**: **Mặc dù** không bị nguy hiểm đến tính mạng nhưng nạn nhân vẫn chưa thể làm việc với cơ quan điều tra. (***Although** the victim's life is not in danger, the victim is still unable to work with the investigative agency.*) **Explanation**: Causal, Although clauses, etc., can be used to create the entailment hypothesis sentence with the premise sentence. This is a case of Others.	2%

Table 11: Examples of entailment rules for creating premise (**P**) - hypothesis (**H**) pairs. Simply, we only mention one rule to be applied in each example.