ViHealthNLI: A Dataset for Vietnamese Natural Language Inference in Healthcare

Huyen Nguyen[1], **The-Quyen Ngo**[1], **Thanh-Ha Do**[1], **Tuan-Anh Hoang**[2,*]
[1]University of Science, Vietnam National University Hanoi, [2]RMIT University Vietnam
{huyenntm, ngoquyenbg}@hus.edu.vn, hadt_tct@vnu.edu.vn, anh.hoang62@rmit.edu.vn

Abstract

This paper introduces ViHealthNLI, a large dataset for the natural language inference problem for Vietnamese. Unlike the similar Vietnamese datasets, ours is specific to the healthcare domain. We conducted an exploratory analysis to characterize the dataset and evaluated the state-of-the-art methods on the dataset. Our findings indicate that the dataset poses significant challenges while also holding promise for further advanced research and the creation of practical applications.

Keywords: Natural language inference, Vietnamese, Healthcare

1. Introduction

The natural language inference (NLI) problem requires us to determine the semantic relationship between a pair of input sentences - a *premise* and a *hypothesis*. This relationship can be either *entailment* (if the hypothesis can be inferred from the premise), *contradiction* (if the negation of the hypothesis can be inferred from the premise), or *neutral* (for all the other cases). Recent studies have highlighted the critical role of NLI in many vital applications (Yang et al., 2019; Glockner et al., 2024), particularly in the healthcare domain (Sarrouti et al., 2021; Arana-Catania et al., 2022). Over the past decade, this problem has attracted numerous studies (Storks et al., 2019; Gubelmann et al., 2023). Thanks to the creation of large scale datasets in English(Bowman et al., 2015; Williams et al., 2018), researchers have proposed a multiple models for the problem with impressive performance [1]. However, their performance for other languages, including Vietnamese, still needs to improve. This decline in the models' performance is primarily due to the lack of appropriate datasets.

Despite having a large number of speakers and a rapidly growing demand for language technologies[2], Vietnamese is still a low-resource language. Particularly for NLI, to the best of our knowledge, ViNLI (Huynh et al., 2022) is the only existing dataset for Vietnamese. However, this dataset is open-domain, making it unsuitable for use in certain specific domains (Bauer et al., 2021).

In this work, we aim to address the above issues by constructing a novel domain-specific dataset for NLI for Vietnamese. Our work is also motivated by the recent campaigns[3] and the emerging need for tools for assessing health information in Vietnam[4]. Hence, we include in the dataset sentences about healthcare topics and events, and name it ViHealthNLI. We have performed an initial analysis to explore the subjects discussed in the dataset. We have also examined the effectiveness of several state-of-the-art methods on the dataset. The findings demonstrate that our dataset poses significant novelty, and suggests promising applications.

2. Related Work

Multiple datasets were created to facilitate the development of advanced methods for the NLI problem. The first ones, quite limited in size, were introduced in RTE challenges (Dagan et al., 2006, 2010). Larger datasets were then constructed and publicly released. The notable are the SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). Recently, more comprehensive datasets have been constructed to improve the NLI models further (Nie et al., 2019, 2020; Conneau et al., 2020b; Parrish et al., 2021; Liu et al., 2022). These datasets are, however, only in English and open-domain. There are also several datasets tailored for other languages. Hu et al. introduced the OCNLI dataset for Chinese (Hu et al., 2020), Mahendra et al. developed IndoNLI for Indonesian (Mahendra et al., 2021), and Yanaka et al. presented JaNLI dataset for Japanese (Yanaka and Mineshima, 2021). Specifically, Huynh et al. curated ViNLI, a dataset focusing on Vietnamese (Huynh et al., 2022). These datasets, however, primarily serve open-domain purposes, lacking specific domain constraints or focuses.

Unlike the work above, we focus on constructing a large NLI benchmark dataset specifically tailored

* Corresponding author

[1]https://paperswithcode.com/task/natural-language-inference

[2]https://www.statista.com/forecasts/1147008/internet-users-in-vietnam

[3]https://en.vietnamplus.vn/campaign-seeks-to-prevent-fake-news-create-healthier-cyberculture/269457.vnp

[4]https://indochina-research.com/4-out-of-10-vietnamese-youth-are-exposed-to-fake-news/

SIGUL2024 Workshop, pages 404–409
21-22 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

for Vietnamese and the healthcare domain. Moreover, we rigorously oversee the data compilation procedure to minimize any annotation artifacts and bias that found present in the current datasets (Gururangan et al., 2018).

3. Data Collection

We use the well-established technique in previous studies to construct datasets. This technique involves the following primary phases:

- Phase 1: Choosing the first sentence, known as the "premise," from a text source about healthcare, followed by
- Tasking human annotators with crafting the subsequent sentence, the "hypothesis," which either logically follows, opposes, or remains impartially related to the chosen sentence.

In phase 1, following the previous work that constructed the ViNLI dataset, we also use news articles as the source for choosing the premise sentences. That type of source is also used to serve our objective: We would like to employ the constructed dataset to develop tools for information verification in the news. To do so, we first crawled news articles from reputable and highly popular online news agencies in Vietnam, such as VnExpress[5], Dan Tri[6], Tuoi Tre[7], and others. We only crawl articles published under the *Health* category of the agencies to focus on healthcare-related topics. In total, we have crawled more than 10 thousand articles published in the last three years. Next, we selected the first sentences from the those articles as potential premise sentences. These sentences were chosen due to their semantic conciseness: Their meaning can be comprehensively understood based solely on their wording. We then exclude sentences with fewer than ten words or end with exclamation marks or question marks since these sentences often do not provide factual information.

In phase 2, we recruited a large group of undergraduate students as annotators to compile the hypothesis sentences. To increase the linguistic diversity of the dataset, we selected students from different majors, including science, technology, business and economy-related studies, and art. Additionally, in order to guarantee the annotators possess adequate language skills, we exclusively accepted individuals meeting two criteria: (1) being native speakers of Vietnamese and (2) having reached at least their third year of study in their program. Altogether, our team comprises over 30 annotators.

[5]https://vnexpress.net/

[6]https://dantri.com.vn/

[7]https://tuoitre.vn/

We randomly distributed the premise sentences among the annotators. Each annotator was tasked with generating three additional sentences in Vietnamese for each premise sentence, aiming to convey, respectively, semantic entailment, contradiction, or neutrality with the premise sentence. We supplied the annotators with the following guidelines for constructing each hypothesis sentence.

- **Entailment**: Create a sentence that either (i) implies or restates the key point(s) in the premise sentence by employing synonymous terms and/or (ii) expands upon or clarifies the point(s) while altering the sentence structure.
- **Contradiction**: Create a new sentence that either (i) refutes (one of) the main idea(s) in the premise sentence by using opposite terms or (ii) restates the primary actions/statements/opinions/etc. mentioned in the translated premise sentence using synonyms but with different subjects and/or objects, along with making any necessary structural adjustments for linguistic fluency.
- **Neutral**: to compose a new sentence that mentions one or more subject(s) of the translated premise sentence but discusses aspects not mentioned in that sentence.

Moreover, we implemented several pilot sessions to train the annotators. In each session, annotators were tasked with working on a few premise sentences and refining their hypothesis sentences with the help of senior researchers. The refinement focuses on avoiding direct affirmations or negations and discouraging mere replication of premise sentences in composing the hypotheses. As highlighted in (Gururangan et al., 2018), this refinement is necessary to minimize annotation artifacts and biases in data construction. Following the training, we allocated the premise sentences to annotators in sizable groups. Two annotators then worked on each group: one compiled the hypothesis sentences, and the other revised the sentences based on the aforementioned revision guidelines.

4. Data Validation

To ensure the reliability of our dataset, we conducted data verification by selecting 500 pairs of (premise, hypothesis) sentences randomly for validation. These pairs were relabeled by 3 to 5 senior researchers without knowledge of the original annotators or labels. Additionally, we randomized the order of sentences within pairs and the presentation order of pairs to senior researchers. Remarkably, 98.2% of pairs received unanimous labeling from senior researchers, leading to high agreement. Utilizing the majority voting method, we

Table 1: Basic statistics of the ViNLI and ViHealthNLI datasets.

Statistic	ViNLI	ViHealth
#pairs	22,801	18,989
#entailment pairs	7,583	6,398
#contradiction pairs	7,595	6,333
#neutral pairs	7,623	6,258
average #words in premise sentences	28.6	26.8
average #words in hypothesis sentences:		
- entailment sentences	19.5	25.4
- contradiction sentences	18.3	22.2
- neutral sentences	21.7	22.3

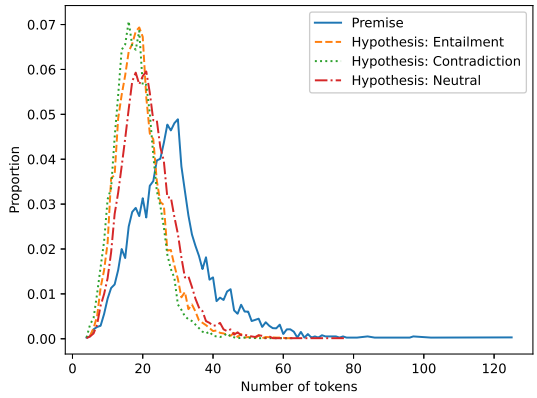

Figure 1: Length distribution of sentences in the ViHealthNLI dataset.

unified the researchers' labels for each pair, resulting in 97.8% agreement between annotators and senior researchers. These findings underscore the quality and trustworthiness of our dataset.

5. Descriptive Analysis

First, in Table 1, we show some basic descriptive statistics of the ViHealthNLI dataset. The table also presents comparative statistics from the ViNLI dataset[8], the only publicly available dataset for Vietnamese NLI. The table clearly shows that while the ViNLI dataset is slightly larger, the ViHealthNLI dataset is slightly more comprehensive, as their hypothesis sentences are significantly longer.

Next, in continuation of prior research, we delved deeper into the length of the sentences and the linguistic overlapping between the premise and the hypothesis sentences in our dataset. We show in Figure 1 the distributions of the length, and in Figure 2 the distributions of the overlappings. Here, a sentence's length is measured by the number of its tokens, and the overlapping between two sentences is measured by the Jaccard coefficient between the set of sentences' uncased tokens. The figures indicate that both the length and the overlapping adhere to long-tailed normal distributions, implying the complexity of the dataset.

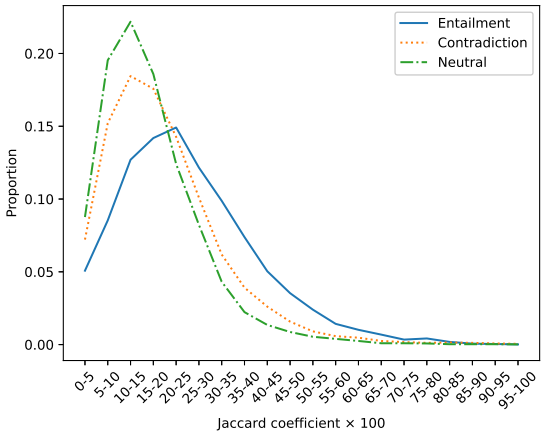

Figure 2: Distribution of the overlapping between the premise sentence and the hypothesis sentence in ViHealthNLI dataset.

Lastly, we performed a topical examination to gain insight into the subjects covered within our dataset. We utilized the LDA approach (Blei et al., 2003), setting the number of hidden topics to 5 after a thorough exploration involving various values, considering the balance between the model's likelihood and the coherence of the identified topics (Wallach et al., 2009). In Table 2, we show the proportion and top 10 most representative words for each obtained topic. The table also shows the topics' label, which is manually assigned based on the topics' most representative words and sentences.

6. Annotation Artifact Examination

Like previous studies, we examined annotation artifacts within our dataset by predicting the labels of hypothesis sentences without considering the premise sentences. We used *Na"ive Bayes*[9] and fasttext models[10] for the task and implemented 5-fold cross-validation. We show in Table 3 the aggregated performance of the models across the folds. Additionally, we include in the table the performance of the identical experiments on the ViNLI dataset. The table indicates that our ViHealthNLI dataset exhibits a slightly higher occurrence of annotation artifacts than the ViNLI dataset. This discrepancy is anticipated since the ViNLI dataset en-

[8]We exclude from VinNLI dataset pair of *Other* category to make it consistent with other datasets

[9]https://nlp.stanford.edu/IR-book/pdf/13bayes.pdf

[10]https://fasttext.cc/docs/en/supervised-tutorial.html

Table 2: Topics obtained from ViHealthNLI dataset: the label assigned to each topic is manually determined based on examining the topic's top words and top sentences.

Topic	Proportion	Label	Top words(translated into English)
1	17.1%	Cardiovascular health	pain, blood, disease, joint, inflammation, doctor, surgery, patient, heart, hospital
2	23.2%	Fertility and children	child, skin, baby, mother, pregnancy, help, birth, health, women, pregnant
3	17.9%	Nutrition	health, help, weight, nutrition, food, body, substances, benefits, regimen, drink
4	18.9%	Covid-19	covid-19, medical, case, hospital, patient, vaccine, epidemic, province, disease, Vietnam
5	22.9%	Cancer	disease, cancer, treatment, inflammation, infection, symptoms, medicine, risk, help, danger

Table 3: The average micro F1 score of hypothesis sentence classifiers.

Model	ViNLI	ViHealthNLI
Naïve Bayes	0.466	0.495
fasttext	0.492	0.531

compasses a broader range of domains, whereas our ViHealthNLI dataset is specific to a particular domain. It is worth noting that the classifier's performance on our dataset is notably lower compared to similar results on existing datasets (Gururangan et al., 2018), suggesting a significant reduction in annotation artifact issues in our dataset.

7. Experiment

We first examine the effectiveness of the state-of-the-art pre-trained model on our datasets. We used a version of the DeBERTaV3 model that was initially trained on a huge multilingual dataset and then fine-tuned on MNLI and XNLI datasets[11]. This model obtains an accuracy of only 82.6% on ViHealthNLI, significantly lower than its performance on English datasets[12], highlighting the difficulty in performing cross-lingual transfer learning on our dataset.

Next, in line with previous research, we investigate the efficiency of transformer-based classification models, which have demonstrated superiority in various natural language comprehension tasks, including NLI, as highlighted in recent studies (Min et al., 2023). Specifically, we used **phobert-based**[13] and **phobert-large**[14] – as they are the most performant BERT for Vietnamese (Nguyen and Nguyen, 2020); **xmlr-base**[15] and **xmlr-large**[16] – the pre-trained XLM-RoBERTa models (Conneau et al., 2020a); and **deberta-v3-base**[17] and **deberta-v3-large**[18] – the pre-trained DeBERTaV3 models (He et al., 2022). We conducted a 5-fold cross-validation for each model utilizing Hugging Face's library, [19] employing hyper-parameter configurations include *learning-rate* = 10^{-5}, *batch-size* = 32, *number-epochs* = 5. Table 4 shows the models' average accuracy across the folds. It is evident from the table that the models achieve comparable results to the current state-of-the-arts, implying that our dataset presents a difficulty while also providing significant prospects for future sophisticated research and the creation of practical applications.

Table 4: Performances of transformer-based models on ViHealthNLI dataset.

Model(s)	Avg. Accuracy
phobert-base	0.900
phobert-large	0.914
xlmr-base	0.877
xlmr-large	0.913
deberta-v3-base	0.809
deberta-v3-large	0.862

Lastly, we performed a cross-dataset evaluation to obtain a qualitative comparison between our dataset and ViNLI dataset. We trained a phobert-large-based classification model on one dataset and tested it on the other. For the train on ViNLI and test on ViHealthNLI, we obtained an accuracy of 85.4%, and for the train on ViHealthNLI and test on ViNLI, we obtained an accuracy of 64.5%. These results clearly imply the significant qualita-

[11] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-mnli-xnli

[12] https://paperswithcode.com/paper/deberta-decoding-enhanced-bert-with

[13] https://huggingface.co/vinai/phobert-base

[14] https://huggingface.co/vinai/phobert-large

[15] https://huggingface.co/xlm-roberta-base

[16] https://huggingface.co/xlm-roberta-large

[17] https://huggingface.co/microsoft/deberta-v3-base

[18] https://huggingface.co/microsoft/deberta-v3-large

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

tive difference between the two datasets.

8. Conclusion

We have provided a large, novel dataset for the NLI problem in Vietnamese that is specific to the healthcare domain. We have also conducted several experiments to get insight from the dataset and examine the state-of-the-art models on it. The findings suggest that the dataset has the potential to explore more domain-specific research as well as practical applications, such as in combatting misinformation (Yang et al., 2019).

9. Acknowledgement

This work is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14

10. Bibliographical References

Miguel Arana-Catania, Elena Kochkina, Arkaitz Zubiaga, Maria Liakata, Robert Procter, and Yulan He. 2022. Natural language inference with self-attention for veracity assessment of pandemic claims. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1496–1511.

Lisa Bauer, Lingjia Deng, and Mohit Bansal. 2021. Ernie-nli: Analyzing the impact of domain-specific external knowledge on enhanced representations for nli. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 58–69, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

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.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R Bowman. 2020b. Xnli: Evaluating cross-lingual sentence representations. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2475–2485. Association for Computational Linguistics.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. Ambifc: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. 2023. Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks. *Journal of Logic, Language and Information*, pages 1–28.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 107–112. Association for Computational Linguistics (ACL).

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. Indonli: A natural language inference dataset for indonesian. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 10511–10527. Association for Computational Linguistics (ACL).

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

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.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

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*. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 1112–1122. Association for Computational Linguistics (ACL).

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the generalization capacity of pre-trained language models through japanese adversarial natural language inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349.

K-C Yang, T Niven, and Hung-Yu Kao. 2019. 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)*.