Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19671-1

LMCK: pre-trained language models enhanced with contextual knowledge for Vietnamese natural language inference

Ngan Luu-Thuy Nguyen[1,2] · Khoa Thi-Kim Phan[1,2] · Tin Van Huynh[1,2] · Kiet Van Nguyen[1,2]

Received: 18 October 2023 / Revised: 14 April 2024 / Accepted: 10 June 2024

Abstract

Natural Language Inference (NLI) has gathered significant attention in recent years due to its application. However, to apply to other downstream tasks, the NLI task should be extended its boundaries by adopting prominent approaches such as looking beyond the sentence level, taking advantage of linguistic phenomena, or eventually providing world knowledge. Therefore, numerous works have been conducted in recent years on various benchmark datasets. In this work, we proposed LMCK, a natural language inference mechanism utilizing pre-trained language models and context-based external knowledge applied to the premise of the Vietnamese dataset. We also investigate popular pre-trained language models for the NLI task at the passage level and employ different information retrieval models. Our findings show that: (1) A longer premise is indeed a primary determinant for improving performance on the NLI task; nevertheless, the significance lies more in the content within the premise; (2) We observe in this task the encoders give better results than the encoder-decoder; (3) Our approach successfully achieves state-of-the-art performance on the benchmark dataset ViNLI with 4 classes.

Keywords Natural language inference · Information retrieval · Context-based external knowledge · Pre-trained language models

✉ Kiet Van Nguyen
kietnv@uit.edu.vn

Ngan Luu-Thuy Nguyen
ngannlt@uit.edu.vn

Khoa Thi-Kim Phan
khoaptk@uit.edu.vn

Tin Van Huynh
tinhv@uit.edu.vn

[1] University of Information Technology, Ho Chi Minh City, Vietnam

[2] Vietnam National University, Ho Chi Minh City, Vietnam

Published online: 22 June 2024

1 Introduction

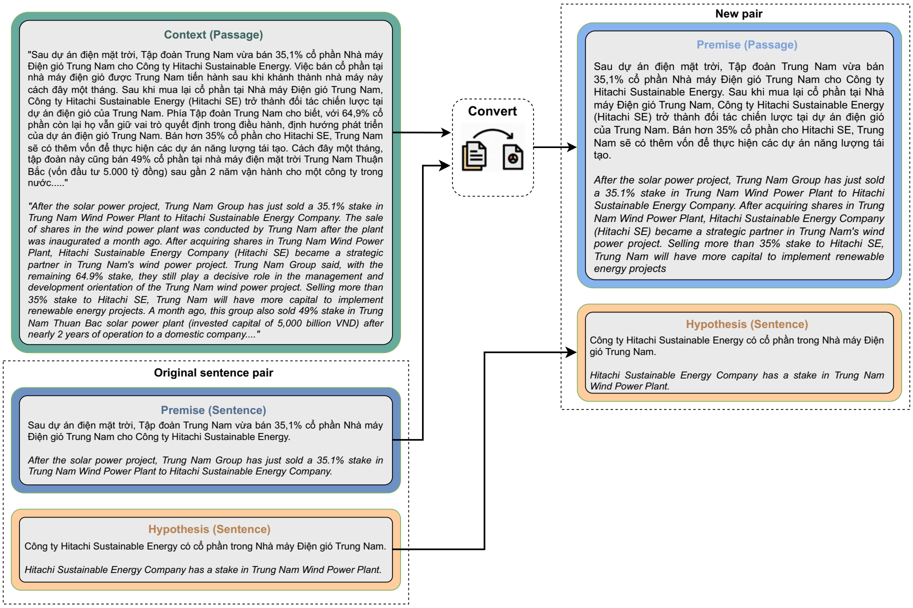
Natural Language Processing (NLP) is an indispensable contributor to the great rise of Artificial Intelligence (AI) around the world. One of the NLP downstream tasks having a plethora of practical applications is Recognizing Textual Entailment (RTE), also called Natural Language Inference (NLI). With the goal of determining whether a hypothesis of natural language h can be inferred from a given premise p, NLI is often treated as a classification problem: given two inputs - hypothesis and premise - the problem is to classify the relationship between them into one of three classes: 'entailment', 'contradiction' or 'neutral'. Besides, it is evident that the majority of the forms of meaningfulness in language can be considered as a form of entailment, contradiction, and neutrality in context [1, 2]. Hence, NLI has played a crucial role in advance of NLP's downstream applications such as Question Answering, Text Summarization, and Machine Reading Comprehension.

Currently, there are many works that promote the development of this field, involving publishing high-quality NLI datasets, as well as improving NLI models to be comparable to the level of human beings. In particular, a plethora of large-scale datasets for the task NLI in various languages or domains has been published such as SNLI [3], MultiNLI [4], and ViNLI [5]. On the other hand, architecture-oriented branches (i.e., Transformer-based language models) such as XLM-R [6], InfoXLM [7], PhoBERT [8] and mBART [9] have been working well on this task. Moreover, these models have outperformed the performance of the non-expert human when being fine-tuned and evaluated on different benchmark datasets.

However, applying NLI development to other downstream NLP tasks effectively still need a lot of attempt by the NLP community. One of these factors that were shown to affect the performance of the models is the length of the premise [10]. Most recent works have been done to address the task at the sentence level, which might be a lack of contextual information. As a result, although still achieving competitive results, these models demonstrated that they are not good at performing inference over longer text, which is a main feature of the NLP downstream tasks [10]. As indicated in [11], the inference is made based on contextual information and a collection of facts. Deducing and then connecting hidden facts from a given context is an essential part of human language understanding, involving many steps and much information. Hence, only using the information from a sentence might not be enough to address sufficiently the NLP downstream tasks which require processing long text. Therefore, many works investigating NLI tasks at the passage level have received much attention [10, 12, 13].

In the Vietnamese domain, to our knowledge, the monolingual dataset for NLI task is quite rare, just including ViNLI [5] and a bilingual dataset Vietnamese-English NLI [14]. In addition, these Vietnamese datasets are on the sentence level. Therefore, to conduct experiments to investigate whether a longer premise can improve the performance of models, we leveraged the contexts that are additionally provided in the ViNLI dataset to generate a long-premise ViNLI dataset, as shown in Fig. 1. Compared to other benchmark datasets, ViNLI [5] was designed into 4 labels (ENTAILMENT, CONTRADICTION, NEUTRAL, and OTHER) instead of three (ENTAILMENT, CONTRADICTION, and NEUTRAL) due to some certain circumstances in real-life scenarios.

In this paper, we restrict our focus in solving the NLI task to the “entailment” class, as it plays an important role in downstream tasks such as Question Answering; for the other three classes, we remain unchanged. Moreover, we not only emphasize the need for a long-premise NLI dataset, but we also pay attention to how valuable information is in the premise. Specifically, we develop a framework named LMCK that uses pre-trained language models



Context (Passage)

"Sau dự án điện mặt trời, Tập đoàn Trung Nam vừa bán 35,1% cổ phần Nhà máy Điện gió Trung Nam cho Công ty Hitachi Sustainable Energy. Việc bán cổ phần tại nhà máy điện gió được Trung Nam tiến hành sau khi khánh thành nhà máy này cách đây một tháng. Sau khi mua lại cổ phần tại Nhà máy Điện gió Trung Nam, Công ty Hitachi Sustainable Energy (Hitachi SE) trở thành đối tác chiến lược tại dự án điện gió của Trung Nam. 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. Bán hơn 35% cổ phần cho Hitachi SE, Trung Nam sẽ có thêm vốn để thực hiện các dự án năng lượng tái tạo. Cách đây một tháng, tập đoàn này cũng bán 49% cổ phần tại nhà máy điện mặt trời Trung Nam Thuận Bắc (vốn đầu tư 5.000 tỷ đồng) sau gần 2 năm vận hành cho một công ty trong nước....."

"After the solar power project, Trung Nam Group has just sold a 35.1% stake in Trung Nam Wind Power Plant to Hitachi Sustainable Energy Company. The sale of shares in the wind power plant was conducted by Trung Nam after the plant was inaugurated a month ago. After acquiring shares in Trung Nam Wind Power Plant, Hitachi Sustainable Energy Company (Hitachi SE) became a strategic partner in Trung Nam's wind power project. Trung Nam Group said, 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. Selling more than 35% stake to Hitachi SE, Trung Nam will have more capital to implement renewable energy projects. A month ago, this group also sold 49% stake in Trung Nam Thuan Bac solar power plant (invested capital of 5,000 billion VND) after nearly 2 years of operation to a domestic company...."

Original sentence pair

Premise (Sentence)

Sau dự án điện mặt trời, Tập đoàn Trung Nam vừa bán 35,1% cổ phần Nhà máy Điện gió Trung Nam cho Công ty Hitachi Sustainable Energy.

After the solar power project, Trung Nam Group has just sold a 35.1% stake in Trung Nam Wind Power Plant to Hitachi Sustainable Energy Company.

Hypothesis (Sentence)

Công ty Hitachi Sustainable Energy có cổ phần trong Nhà máy Điện gió Trung Nam.

Hitachi Sustainable Energy Company has a stake in Trung Nam Wind Power Plant.

New pair

Premise (Passage)

Sau dự án điện mặt trời, Tập đoàn Trung Nam vừa bán 35,1% cổ phần Nhà máy Điện gió Trung Nam cho Công ty Hitachi Sustainable Energy. Sau khi mua lại cổ phần tại Nhà máy Điện gió Trung Nam, Công ty Hitachi Sustainable Energy (Hitachi SE) trở thành đối tác chiến lược tại dự án điện gió của Trung Nam. Bán hơn 35% cổ phần cho Hitachi SE, Trung Nam sẽ có thêm vốn để thực hiện các dự án năng lượng tái tạo.

After the solar power project, Trung Nam Group has just sold a 35.1% stake in Trung Nam Wind Power Plant to Hitachi Sustainable Energy Company. After acquiring shares in Trung Nam Wind Power Plant, Hitachi Sustainable Energy Company (Hitachi SE) became a strategic partner in Trung Nam's wind power project. Selling more than 35% stake to Hitachi SE, Trung Nam will have more capital to implement renewable energy projects

Hypothesis (Sentence)

Công ty Hitachi Sustainable Energy có cổ phần trong Nhà máy Điện gió Trung Nam.

Hitachi Sustainable Energy Company has a stake in Trung Nam Wind Power Plant.

Fig. 1 Transformation of the original sentence-level premise ViNLI dataset into the longer-premise ViNLI dataset based on the provided context

collaborated with context-based external knowledge generated by combining our rules with information retrieval models such as BM25 [15], TF-IDF [16], Sentence-Bert [17], SXLM-R [18] for our experiments. We also experiment with two types of pre-trained models on the NLI task: the encoder (XLM-R, PhoBERT, InfoXLM), and the encoder-decoder (mBART) on our converted long-premise NLI dataset. Our investigation demonstrates that besides longer premises, context-based external knowledge is an important factor for better performance on NLI task. For this task, the results display that encoders are better than the encoder-decoder. Most importantly, our approach achieves state-of-the-art performance on the ViNLI dataset.

The rest of the paper is organized as follows. In Section 2, we provide an overview of previous works about the Natural Language Inference task and Context-based external knowledge. Section 3 describes the methodology which is used for experiments in this paper. Then, we present the whole experiment, including the dataset, experimental settings, and our results in Section 4. Finally, Section 5 presents the conclusion and future works.

2 Related works

We consider previous works in the areas of both Natural Language Inference Section 2.1 and Context-based external knowledge Section 2.2.

2.1 Natural language inference

Since 2005, the NLP community has witnessed a significantly growing popularity of the task Recognizing Textual Entailment(RTE), which is now known as Natural Language Inference

(NLI) due to the emergence of the PASCAL Recognizing Textual Entailment (RTE) challenges [19]. The key to the popularity is that the RTE task works as a system in which to determine the relationship between two given text fragments by employing different techniques used in NLP applications to address semantic inference which is a prominent issue shared by many NLP applications. 2 years later, on the third RTE challenge [20], a limited number of longer texts, i.e. up to a paragraph in length, were introduced to make the challenge more oriented to realistic scenarios, which is one of the most inspirational works to later works related to RTE and its applications. After that, these further RTE challenges such as RTE-5 [21], RTE-6 [22], RTE-7 [23] required communities to mainly apply RTE systems to specific application settings. In particular, all three challenges (RTE-5, RTE-6, RTE-7) are situated in the Summarization application setting.

Recently, with the challenge of more comprehensive scenarios, there has been a plethora of work improving both datasets and techniques. On one hand, the most well-known NLI benchmarks include the Standford Natural Inference (SNLI) dataset [3]; and the expanded Multi Genre NLI corpus(MultiNLI) [4] attempting to tackle the limitations of SNLI. Specifically, the dataset introduced various genre labels for each sentence pair to concentrate on domain adaption. Besides, there are several task-specific NLI datasets, consisting of Question-answering NLI (QNLI) [24], SciTail [25], Dialogue NLI [26], and Vietnamese-English NLI [14]. In addition, there is also various monolingual NLI dataset, including OCNLI [27] for Chinese, IndoNLI [28] for Indonesian, SICKNL [29] for Dutch, and ViNLI [5] for Vietnamese. However, all the above datasets are either on the sentence level or do not consider the relationships that infer from more than sentences.

Therefore, NLI datasets with longer text have been built as a necessity to address inferences in real-life situations. In 2014, the Approximate Textual Entailment (ATE) dataset used in the field of Image Captioning [30] was created based on *FLICKR30k*. Each item includes a premise set of four captions and a short phrase as the hypothesis. Similarly, the Multiple Premise Entailment (MPE) datasets [31] was proposed as a challenging task in which each hypothesis sentence is paired with an unordered set of written premise sentences that demonstrate the same event from *FLICKR30k*. Regarding the field NLI, Adversarial NLI [32] is a novel human-and-model-in-the-loop dataset in which longer contexts are considered in the premise. The ConTRol [13] is a dataset for contextual reasoning over long texts. Compared to Adversarial NLI, the context of ConTRol is much longer and described under multiple paragraphs, while Adversarial NLI has only single-paragraph contexts. As inspired by these above works, to investigate the potential of longer-premise in dealing with the Vietnamese NLI task, we leverage the contexts which are additionally provided in the ViNLI dataset and convert the ViNLI dataset from single-sentence premise into multiple-sentence premise (i.e. from sentence-level to passage-level).

On the other hand, due to the increasing growth of large-scale NLI datasets, deep learning models such as RNN [33], BiLSTM [34], and ESIM [35] have passed beyond traditional machine learning models (Skip-gram, CBOW [36]). However, in recent years, the advent of transformer architecture [37] completely changed how researchers deal with the NLI task and its applications. In particular, numerous models have proposed and given significant performances by employing both the architecture of the encoder including BERT [38], XLM-R [6], and InfoXLM [7], and that of the encoder-decoder consisting of BART [39], t5 [40]. Also, for Vietnamese transformer-based models, PhoBERT [8] and ViT5 [41] have done positive results for the Vietnamese domain.

Although the effectiveness of transformer-based models in the NLI task is significant, in this work, we demonstrate that the performance of NLI models that use pre-trained models can be augmented with context-based external knowledge.

As far as Vietnamese NLI is concerned, in the Vietnamese NLP community, NLI is an area that has recently been a new research subject. Therefore, there hasn't been a lot of work yet in this field. The advent of [14] as a shared task in VLSP[1] has drawn more Vietnamese researchers' attention. With great effort, several outstanding works [5, 42–45] were proposed. In particular, the studies [42, 43] are works in the shared task [14]. While [42] utilized pre-trained Multilingual Language Models, [43] employed data augmentation to deal with Vietnamese and English-Vietnamese Textual Entailment tasks. [5] made a major contribution to the Vietnamese NLP community due to the creation of the first monolingual Vietnamese NLI dataset - ViNLI. [44] proposed a method to build a Vietnamese dataset for training Vietnamese inference models that work on native Vietnamese texts. [45] presented an experiment combining semantic word representation through the SRL task with context representation of BERT relative models for the NLI problem. Despite many attempts, there is still no work using context-based external knowledge to enhance the performance of models in the Vietnamese NLI dataset.

2.2 Context-based external knowledge

Utilizing context-based external knowledge has shown improvement in performance on many NLP downstream tasks [13, 46–51]. There are two main approaches utilizing context-based external knowledge, including graph-based and information retrieval-based approaches.

For graph-based external knowledge in the field of Natural Language Inference (NLI), there are a lot of attempts such as [47, 52, 53]. In particular, Wang et al. [47] presented a combination of techniques on text, graph, and text-and-graph-based models that can leverage external knowledge to improve performance on the NLI problem. Chen et al. [52] developed a model with WordNet-based co-attention that uses five engineered features from WordNet for each pair of words from premise and hypothesis. Meanwhile, Pan et al. [53] used an external knowledge source from Knowledge Graphs (KGs) in text-based RTE models by using Personalized PageRank to generate contextual subgraphs with reduced noise and encoding these subgraphs using graph convolutional networks to capture the structural and semantic information in KGs. All in all, most of these works have employed neural networks to represent the triplets of knowledge graphs. These kinds of approaches usually need to train a knowledge-graph embedding beforehand. According to [54], despite the effectiveness, the existing methods for generating knowledge graph embeddings still suffer several severe limitations. In this situation, additional information, such as entity types and relation paths, is ignored, which can further improve the embedding accuracy.

When it comes to information retrieval-based external knowledge, there are two types of representations for retriever: bag-of-word (BOW) based sparse representation [55] and dense representation from neural networks [56]. For the sparse representation, since this method relies on BOW, a rule-based scoring system such as TF-IDF and BM25 is utilized for ranking. This allows for adaptation, to a range of large-scale search scenarios. This method has been widely explored to solve various NLP downstream applications, including Question Answering [57, 58] and Machine Translation [59, 60]. In terms of dense representation based retrieval (DPR) [56], it is the area that has received a lot of attention in recent years. Dense representation is obtained from encoders such as Transformer, trained with task-specific data. It is demonstrated that these methods can yield better recall performance than sparse representation on different tasks. However, DPR cannot process longer documents, usually less than 128 tokens [56].

[1] https://vlsp.org.vn/

In this paper, we focus on getting external knowledge by leveraging information retrieval-based approaches. Therefore, to attain the most suitable retriever for our work, we employ both two types: sparse representation using traditional information retrieval (IR) models such as TF-IDF [16], and BM25 [15], and representation-based retriever using SBERT [17], and the SXLM-R [18].

3 Methodology

Our LMCK system involves a combination of the exploitation of semantic information for the NLI task Section 3.1 and pre-trained language models Section 3.2. In particular, this system includes 3 phases: Context-based Sentence Extraction, Long-premise Generation, and Inference (see Fig 2). As presented in Section 3.1, the Context-based Sentence Extractor, which is the main core of Phase 1, is responsible for extracting external knowledge information from the given context of a document. After that, in phase 2, the most relatable sentences will be added to premise sentences and generate our converted long-premises. For the pre-trained language models Section 3.2 in Phase 3, we use two types of architectures on NLI tasks: the encoder (XLM-R, PhoBERT, InfoXLM), and the encoder-decoder (mBART).

3.1 Context-based sentence extractor

As analyzed in [5], besides depending on the content of the premise, annotators tend to write hypotheses of entailment samples relying on the corresponding premises' situation (i.e. premise's context). Therefore, this might cause difficulties for models in inference. Hence, to facilitate capturing the semantic relations better, we employ the Information Retrieval method together with our rules to get important semantic information.

In this component, the main task requires the retrieval of a proper subset (S_1,S_2, ...,S_n) of each premise's given context from the ViNLI dataset, used for inferring annotators' corresponding hypothesis H, or relating to premise P, or the combination of hypothesis and premise $H + P$. A proper subset results from identifying a subset of statutes for which an entailment system can judge whether the statement H is entailed or not.

In this work, we conduct experiments on two types of information retrieval-based approaches: sparse retriever and representation-based retriever. For sparse retriever, we employ traditional information retrieval (IR) models such as TF-IDF [16], and BM25 [15]. While TF-IDF is a term scoring method using cosine similarity measure, BM25 is a method scoring documents in response to a query. Specifically, TF-IDF and BM25 are respectively displayed in (1), and (2) [61]:

$$TF - IDF(D, Q) = \sum[\sqrt{f(t, D)} * (1 + log(IDF(t)))^2] \tag{1}$$

$$BM25(D, Q) = \sum IDF(q_i) \frac{f(q_i, D) * (k_1 + 1)}{f(q_i, D) + k_1 * (1 - b + b * \frac{|D|}{avgdl})} \tag{2}$$

where D is a document, Q is a query, f(t, D) and f(q_i, D) is t's, q_i's term frequency in document D, |D| is the length of document D in words, avgdl is the average document length in the text collection from which documents are drawn, and IDF is the inverse document frequency. However, there is a slight difference between IDF(t) and IDF(q_i) shown respectively

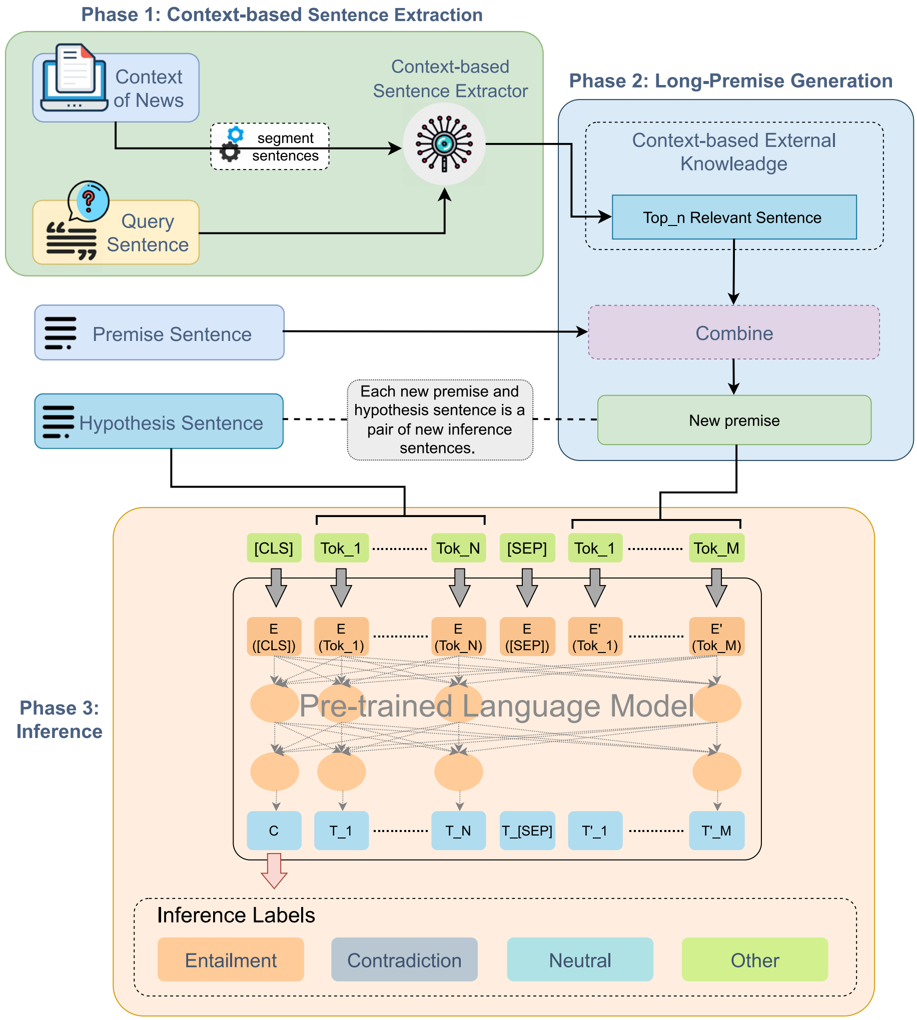


Fig. 2 Overview of large language models enhanced with contextual knowledge (LMCK) system

in (3), and (4). k_1 and b are free parameters. In this work, we set 1.5 for k_1 and 0.75 for b.

$$IDF(t) = log\frac{N}{df(t)} + 1 \tag{3}$$

$$IDF(q_i) = ln(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1) \tag{4}$$

where N is the total number of documents in the collection, df(t) is the number of documents containing t [2], and n(q_i) is the number of documents containing q_i [3].

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

[3] https://en.wikipedia.org/wiki/Okapi_BM25

For representation-based retriever, we use SBERT [17], and SXLM-R [18]. Representation-based retriever, also called Dual-encoder, employs two independent encoders such as BERT [38] to encode the query and the documents respectively, and then estimate their relevance by computing a single similarity score between two representations. In particular, SBERT [17] adopts two independent BERT-based encoders to encode two input sentences, then adds a pooling operation to the output of BERT to derive a fixed-sized sentence embedding. Finally, the relevance score between them is computed by the cosine-similarity. In order to fine-tune BERT, they create siamese and triplet networks [62] to update the weights so that the produced sentence embeddings are semantically meaningful and can be evaluated by cosine-similarity. Compared to SBERT [17], SXLM-R [18] employed two independent XLM-R encoders, and is fine-tuned with Multiple negatives ranking (MNR) loss [63]. The loss function is given by (5):

$$L = -\frac{1}{N} \cdot \frac{1}{K} \cdot \sum_{i=1}^{K} [S(x_i, y_i) - log \sum_{j=1}^{K} e^S(x_i, y_i)] \tag{5}$$

To evaluate and choose the best IR method, an evaluation dataset is created manually to assess the accuracy of these models by 3 well-educated annotators. Firstly, we provide them with the same dataset, including pairs of sentences: premise and hypothesis, and a context stemming from the training set of ViNLI. We require them to read carefully the content of the premise and hypothesis, then check whether we need more contextual information when generating a hypothesis from the premise. If so, annotators will choose the most 3 relevant sentences in the context they think the hypothesis was created based on. We have the most relevant sentence, the second most relevant sentence, and the third most relevant sentence as Top_1, Top_2, and Top_3, respectively. In this process, annotators will work independently. At the end of the process, we only select those samples that all three annotators agree that context is important to writing the hypothesis. As a result, we have three datasets corresponding to three annotators with the same size is 300 samples. Figure 3 shows our evaluation data example.

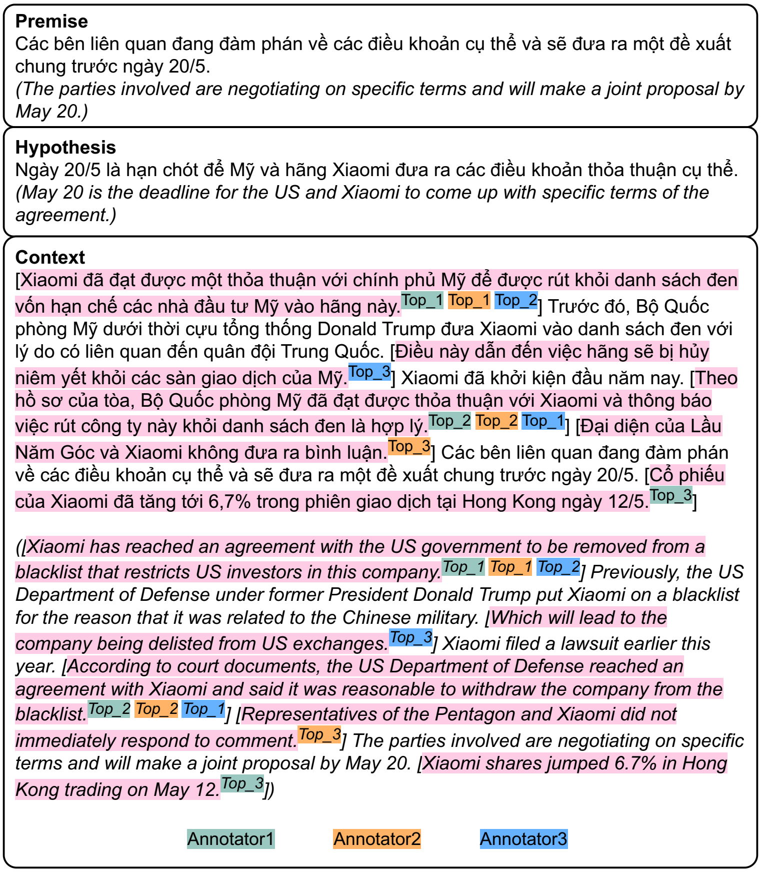
Besides, we design three experiments to evaluate these IR models on the dataset. The difference between the three experiments is the inputs of the IR models, which are described in Fig. 4. After processing inputs into respective embeddings, these IR models calculate the similarity between these embeddings, then return a list of 3 context-based sentences sorted by most relevance.

We use accuracy@3 (i.e. Acc@3) to evaluate the effectiveness of these IR models. For each model, the final accuracy result is the mean of Acc@3 over 3 annotators. Acc@3 is computed as follows:

$$Acc@3 = \frac{\mathrm{X} * 100}{N} \tag{6}$$

where X is the number of predicted sentences appearing in the collective of sentences selected by annotators. N is the total number of annotators in this work.

The results of the evaluation of the IR models are shown in Table 1. We observed that in most models, the pre-trained model SXLM-R [18] gave the highest results in most experiments. Specifically, in Experiment 1, this model achieved 57.22, and in Experiment 2 and Experiment 3, the model attained an accuracy of 55.33 and 59.11, respectively. Therefore, the SXLM-R model is the core of our Information Retrieval component.



Premise
Các bên liên quan đang đàm phán về các điều khoản cụ thể và sẽ đưa ra một đề xuất chung trước ngày 20/5.
(The parties involved are negotiating on specific terms and will make a joint proposal by May 20.)

Hypothesis
Ngày 20/5 là hạn chót để Mỹ và hãng Xiaomi đưa ra các điều khoản thỏa thuận cụ thể.
(May 20 is the deadline for the US and Xiaomi to come up with specific terms of the agreement.)

Context
[Xiaomi đã đạt được một thỏa thuận với chính phủ Mỹ để được rút khỏi danh sách đen vốn hạn chế các nhà đầu tư Mỹ vào hãng này.Top_1 Top_1 Top_2] Trước đó, Bộ Quốc phòng Mỹ dưới thời cựu tổng thống Donald Trump đưa Xiaomi vào danh sách đen với lý do có liên quan đến quân đội Trung Quốc. [Điều này dẫn đến việc hãng sẽ bị hủy niêm yết khỏi các sàn giao dịch của Mỹ.Top_3] Xiaomi đã khởi kiện đầu năm nay. [Theo hồ sơ của tòa, Bộ Quốc phòng Mỹ đã đạt được thỏa thuận với Xiaomi và thông báo việc rút công ty này khỏi danh sách đen là hợp lý.Top_2 Top_2 Top_1] [Đại diện của Lầu Năm Góc và Xiaomi không đưa ra bình luận.Top_3] Các bên liên quan đang đàm phán về các điều khoản cụ thể và sẽ đưa ra một đề xuất chung trước ngày 20/5. [Cổ phiếu của Xiaomi đã tăng tới 6,7% trong phiên giao dịch tại Hong Kong ngày 12/5.Top_3]

([Xiaomi has reached an agreement with the US government to be removed from a blacklist that restricts US investors in this company.Top_1 Top_1 Top_2] Previously, the US Department of Defense under former President Donald Trump put Xiaomi on a blacklist for the reason that it was related to the Chinese military. [Which will lead to the company being delisted from US exchanges.Top_3] Xiaomi filed a lawsuit earlier this year. [According to court documents, the US Department of Defense reached an agreement with Xiaomi and said it was reasonable to withdraw the company from the blacklist.Top_2 Top_2 Top_1] [Representatives of the Pentagon and Xiaomi did not immediately respond to comment.Top_3] The parties involved are negotiating on specific terms and will make a joint proposal by May 20. [Xiaomi shares jumped 6.7% in Hong Kong trading on May 12.Top_3])

Annotator1 Annotator2 Annotator3

Fig. 3 An example of manually generating data to evaluate IR models. In the example, the sentences Top_1, Top_2, and Top_3 are highlighted with green, orange, and blue colors representing the choices of annotator1, annotator2, and annotator3 respectively

Table 1 The results of information retrieval component evaluation according to Acc@3

Experiments's Input	IR Techniques	Acc@3 (%)
Premise & Context	TF-IDF	57.11
	BM25	56.66
	SBERT	**58.77**
	SXLM-R	**57.22**
Hypothesis & Context	TF-IDF	42.11
	BM25	53.44
	SBERT	46.22
	SXLM-R	**55.33**
Premise + Hypothesis & context	TF-IDF	42.88
	BM25	57.66
	SBERT	44.55
	SXLM-R	**59.11**

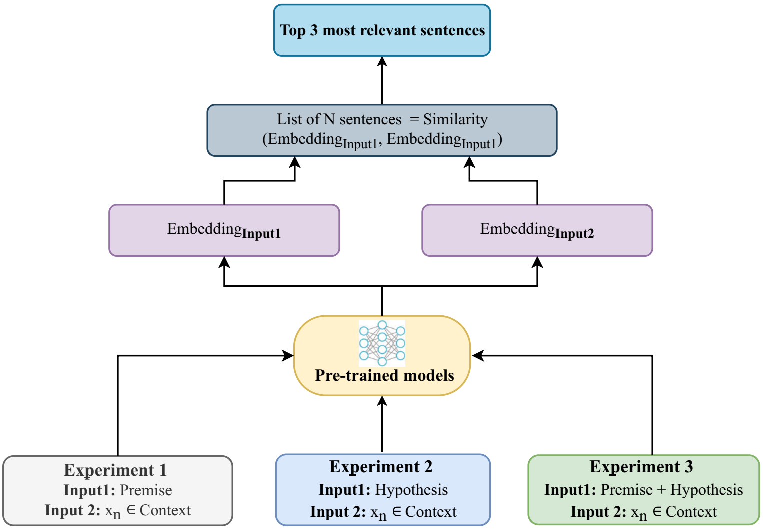

Top 3 most relevant sentences
List of N sentences = Similarity (Embedding$_{Input1}$, Embedding$_{Input1}$)
Embedding$_{Input1}$
Embedding$_{Input2}$
Pre-trained models
Experiment 1
Input1: Premise
Input 2: $x_n \in$ Context
Experiment 2
Input1: Hypothesis
Input 2: $x_n \in$ Context
Experiment 3
Input1: Premise + Hypothesis
Input 2: $x_n \in$ Context

Fig. 4 Three different experiments evaluating these IR models on the dataset

3.2 Pre-trained language models for NLI

To compare with our proposed method, we conduct experiments with several powerful baseline methods using state-of-the-art pre-trained language models.

3.2.1 Pre-trained language models

In this paper, we used four powerful pre-trained language models that are helpful for Vietnamese NLP tasks:

- **PhoBERT** [8] is a monolingual pre-trained model for Vietnamese trained based on RoBERTa [64] with 135M parameters for the base version and 370M for the large version.
- **XLM-R** [6] is an improved version of XLM based on RoBERTa model [64]. XLM-R is trained with a cross-lingual masked language modeling objective on data in 100 languages, including Vietnamese from Common Crawl.
- **InfoXLM** [7] is a multilingual pre-trained model for over 100 languages with a new cross-lingual pre-training task named cross-lingual contrast (XLCO).
- **mBART** [9] is a multilingual encoder-decoder model that is based on BART [39]. mBART is trained with a combination of span masking and sentence shuffling objectives on a subset of 25 languages, including Vietnamese from Common Crawl.

3.2.2 NLI methods using pre-trained language models

The NLI model structures of the encoder and encoder-decoder are illustrated in Fig. 5. For the encoder models (i.e., PhoBERT, XLM-R, and InfoXLM), following [38], given a premise p and a hypothesis h, we concatenate premise-hypothesis pair as a new sequence. However, in this work, due to the new length of the premise, and passage level, compared to other works, we set up a hypothesis and premise respectively instead of the premise, and then hypothesis. Specifically, the input is demonstrated [CLS]+h+[SEP]+p+[SEP], where [CLS]

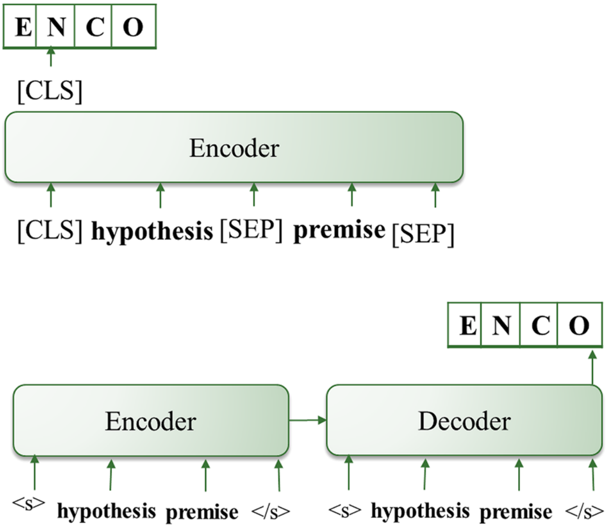

E
N
C
O
[CLS]
Encoder
[CLS] hypothesis [SEP] premise [SEP]
E
N
C
O
Encoder
Decoder
<s> hypothesis premise </s>
<s> hypothesis premise </s>

Fig. 5 The model structure of the encoder and the encoder-decoder. ("E" represents ENTAILMENT, "N" represents NEUTRAL, "C" represents CONTRADICTION, and "O" represents OTHER

and [SEP] are special symbols for the classification token and separator token. After pre-training model encoding, the last layer's hidden representation from the [CLS] token is fed in an MLP+softmax for classification. For the sequence-to-sequence model ((i.e., mBART), we feed the same sequence to both the encoder and the decoder, using the last hidden state for classification. The class corresponding to the highest probability is chosen as the model prediction.

4 Experiments and results

4.1 Dataset and experimental design

After determining the best IR model, we conduct experiment on various types of inputs of models to addressing our research questions. First and foremost, we design 4 different experiments as follows.

- Experiment 1 - **Hypothesis, Context**: Due to the contexts involving premises, and premises' contextual knowledge, we use the given context of a pair of corresponding premises and hypothesis as a premise. The average length of context is 319.9 words.
- Experiment 2 - **Hypothesis, Top(C, P)**: Premise is one of the sentences in a corresponding context that we are received additionally. Therefore, contextual knowledge is supposed to be obtained by applying the best IR SXLM-R [18] with the premise and its context as a premise.
- Experiment 3 - **Hypothesis, Top(C, H)**: As described in [5], the hypothesis was created based on the content of the premise, or the situation of the premise. Thus, we apply the best IR SXLM-R [18] with a hypothesis and its situation (i.e. its context) to obtain contextual knowledge as a premise.

- Experiment 4 - **Hypothesis, Top(C, H+P)**: Due to the relevance of the process of forming premise and hypothesis, we assume we could attain contextual knowledge as a premise by applying the best IR SXLM-R [18] with the combination of hypothesis and premise and its context.

In addition, motivated by how a hypothesis was written, we present a simple rule supporting the generation of better context-based external knowledge as the premise. Our rule is shown in Fig. 6.

As indicated in [5], a hypothesis was created based on the content or situation of the premise (i.e., the context of the premise). Therefore, we strongly believe that we can capture the semantic similarity between context and hypothesis by adapting IR methods. However, after running experiments on the processed dataset, we discovered that the performance of models deteriorates due to information confusion in labels' samples except for entailment. Therefore, we propose the above rule (see Fig. 6) to avoid confusion but achieve our desired improvements, which are used in **Experiment 5**. Specifically, the inputs of the Experiment 5 are listed as follows:

- Input 1: Hypothesis,
- Input 2: Rule Fig. 6 (Premise, Top(C,H))

Most experiments are designed to extract more information from the context to incorporate the premise sentence as the input into the natural language inference model. Whereas the hypothesis statements are kept the same. To observe how these datasets vary in length compared to the ViNLI baseline dataset, we compute the full average length of input 1 (premise + Top_1, Top_2, and Top_3 sentence) of each experiment shown in Table 2. Most of the average length of the experiments' premise is significantly longer than that of the original ViNLI dataset.

Especially with Experiment 1, the length is quite long, with about 330 words. In addition, for each case of extracting 1,2, or 3 sentences in the context to add to the premise sentences of experiments 2,3,4, and 5, the average premise length increases significantly. The average

Fig. 6 Our rule in generating better context-based external knowledge for entailment only

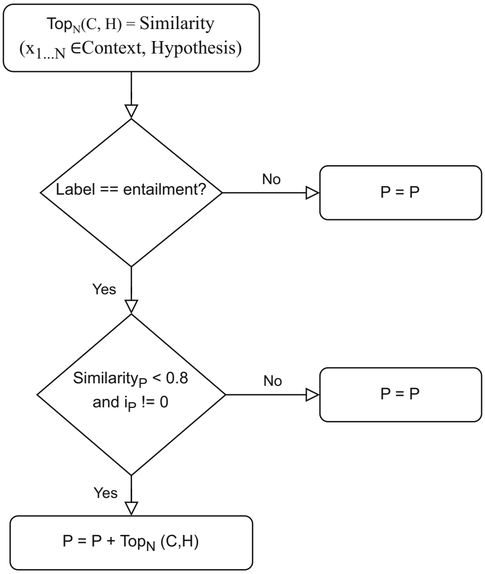

$\text{Top}_N(C, H)$ = Similarity ($x_{1...N}$ ∈Context, Hypothesis)
Label == entailment?
No
P = P
Yes
$\text{Similarity}_P < 0.8$ and i_P != 0
No
P = P
Yes
P = P + Top_N (C,H)

Table 2 The average length of the premise

Premise	Average length (Premise + Top_N)			
		Top_1	Top_2	Top_3
Original ViNLI dataset [5]	24.5	–	–	–
Experiment 1	319.9	–	–	–
Experiment 2	–	50.9	76.5	101.1
Experiment 3	–	25.7	52.3	78.2
Experiment 4	–	24.2	49.1	73.7
Experiment 5	–	28.1	30.8	34.8

Which premise of the "Original ViNLI dataset" is at the sentence level, and the premise of "Experiment1" is an entire context represented as passage level. Meanwhile, the premise of Experiment 2, Experiment 3, Experiment 4, and Experiment 5 is the combination of the premise and its context-based external knowledge attained from the Context-based Sentence Extractor in our system with three cases Top_1, Top_2, and Top_3. Before calculating the average length of the premise, we use the VnCoreNLI tool [65] to segment words for Vietnamese

length statistics are meaningful to us in choosing the max length input parameter of pre-trained transformer models appropriately.

As described above, we can see how the data generated for Experiment 5 differs from the others. The way to generate data for experiments 2, 3, and 4 is always to have context information added to the pairs of inference sentences (There are three cases of +1 sentences, +2 sentences, and +3 sentences) regardless of the difference in labels. Meanwhile, with Experiment 5, we focus on whether it is necessary or not to extract more contextual information to provide sentence pairs of the ENTAILMENT label with context-based external knowledge by setting thresholds and rules in the Context-based Sentence Extractor. In particular, after applying the best IR SXLM-R [18] with a hypothesis and its context, we check whether the label of the sample is Entailment. If yes, we continue to check whether the sample needs more contextual knowledge by using our rules. Therefore, not all sentence pairs of the ENTAILMENT label in the ViNLI dataset need additional contextual information. Figure 7 shows the number of sentence pairs belonging to the ENTAILMENT label in the ViNLI dataset that need and do not need additional contextual information. We found that more than 50% of the sentence pairs of the ENTAILMENT label in the training, development, and test sets of ViNLI need more context. With this considerable amount, we hope that the models trained on the new data can solve the difficult cases of the ENTAILMENT label.

Data generation by adding contextual information to premise sentences, as in our experiments, leads to premise sentence length increasing. While the length of the hypothesis sentence remains the same, the number of words in the hypothesis sentence that do not appear in the premise will also change. We are interested in analyzing this feature of the data because the rate of new words affects the accuracy of the model. Specifically, the research of the author's ViNLI dataset [5] found that the higher the rate of new words, the more difficult it is for the models to predict accurately. Therefore, we focused on analyzing the data of Experiment 5 to observe the new word rate on pairs of sentences of the label ENTAILMENT, as shown in the Table 3. We conduct statistics on all three data creation cases of Experiment 5, which are +Top_1, +Top_2, and +Top_3 sentences with the premise sentence of the

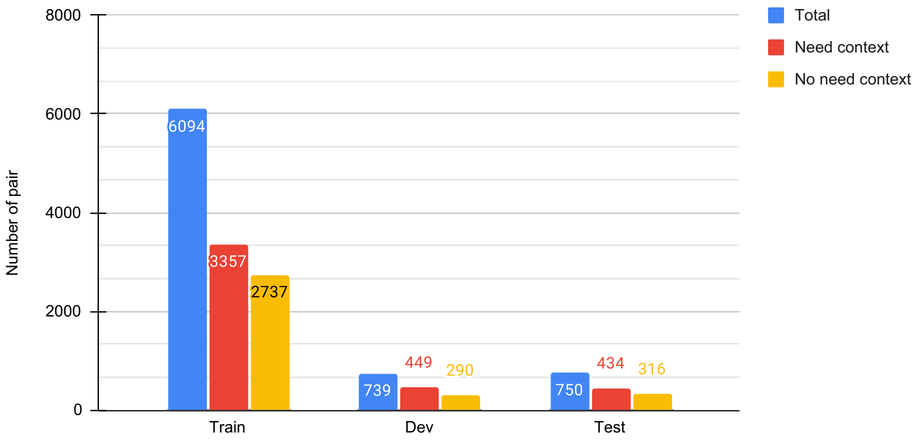


Fig. 7 The number of sentence pairs of ViNLI's ENTAILMENT label needs more context information in Experiment 5

ENTAILMENT label. First, we noticed that the ENTAILMENT label data in all three cases of Experiment 5 has a significantly low rate of new words compared to that on the ENTAILMENT label of the ViNLI dataset. This allows the model to capture better the semantic relationship between premise and hypothesis than the ViNLI dataset. Besides, I also noticed that the new word rate gradually decreased when the premise sentence added Top_1, Top_2, and Top_3, respectively. This can train models to make more accurate predictions when adding necessary context information.

4.2 Experimental settings

In all of our experiments, following the original work on the ViNLI dataset [5], we report the accuracy score as the primary evaluation metric.

As described in Section 4, our approaches depend on pretrained language models such as XLM-R, PhoBERT, mBART, and InfoXLM. Therefore, we use models namely XLM-R$_{large}$, PhoBERT$_{large}$, mBART$_{large}$, InfoXLM$_{large}$ respectively dowloaded from the Hugging Face Library [4].The network's parameters are optimized using the AdamW [66] and a linear learning rate scheduler suggested by the Hugging Face default setup. The hyperparameters that we tuned include the number of epochs, batch size, and learning rate. In particular, we set a batch size of 16 and a learning rate of 1e-5 for all component models. Due to the length of input models, we set the max length to 256 for Top_1, Top_2, and for Top_3, the model is trained on max length 512, where Top_1, Top_2, and Top_3 are the amount of context-based external knowledge representing as sentences. All experiments in this paper are conducted on Google Colab Pro.

4.3 Results and dicussions

According to the Table 4, transformer-based models with the encoder architecture using XLM-R outperform others (PhoBERT, InfoXLM, mBART). Besides, mBART, which is an

[4] https://huggingface.co/transformers/

Table 3 The ratio of new words in the hypothesis sentence compared to the premise sentence on pairs of sentences labeled ENTAILMENT in Experiment 5 compared with the original dataset VINLI

Dataset	New word ate (%)	Part-Of-Speech (%)					
		Noun	Verb	Adjective	Preposition	Adjunct	Other
ViNLI$_{Entailment}$ [5]	46.59	31.45	24.97	6.67	8.39	8.71	19.81
Experiment5_Top_1$_E$	43.71	34.20	22.89	7.07	6.85	7.43	21.56
Experiment5_Top_2$_E$	42.90	34.57	22.64	7.31	6.23	7.25	21.99
Experiment5_Top_3$_E$	40.95	35.17	22.52	7.43	5.67	6.89	22.31

encoder-decoder model almost performs better than PhoBERT, and InfoXLM. According to the table, PhoBERT, a monolingual pre-trained language model for Vietnamese, gives an accuracy of 85.25%, 80.21%, 85.65%, corresponding to Top_1, Top_2, and Top_3. Meanwhile, mBART provides Top_1, Top_2, and Top_3 with respective accuracy of 85.89%, 79.77%, 86.26%. InfoXLM gives an overall accuracy of Top_1, Top_2, Top_3 with 85.02%, 82.81%, and 84.78%, respectively. The top reported performance is given by the XLM-R model, with 89.5% accuracy. Despite our rule only focusing on the "Entailment" label, our approach successfully attains SOTA performance compared to 85.99% in the original.

4.3.1 Model performance on different premise lengths

As mentioned earlier, we designed our experiments beyond the sentence level based on context-based external knowledge, which is represented as multi-sentence. Therefore, after conducting experiments, we compare and contrast the performance of models trained on multiple sentence premise and the single sentence premise in the new dev dataset as in the original to get insight into how context length affects the performance of the transformer-based NLI models. The result is displayed in Fig. 8. When the premise length increases, the model performance drops accordingly. The best model XLM-R drops from 89.23% (Top_1) to

Table 4 Experiment 5 results with the model's input as Top_n(Information Retrieval$_{Entailment}$[Context, Hypothesis]), Hypothesis

Input Top_n(IREntailment[C, H]), H	Model	IR XLM_R + Cosine	
		New Dev	New Test
Top_1, H	XLM_R	89.23	89.50
	PhoBERT	85.71	85.25
	mBART	86.44	85.89
	InfoXLM	85.21	85.02
Top_2, H	XLM_R	89.90	85.38
	PhoBERT	85.94	80.21
	mBART	86.54	79.77
	InfoXLM	88.63	82.81
Top_3, H	XLM_R	89.73	89.13
	PhoBERT	86.21	85.65
	mBART	86.77	86.26
	InfoXLM	84.11	84.78

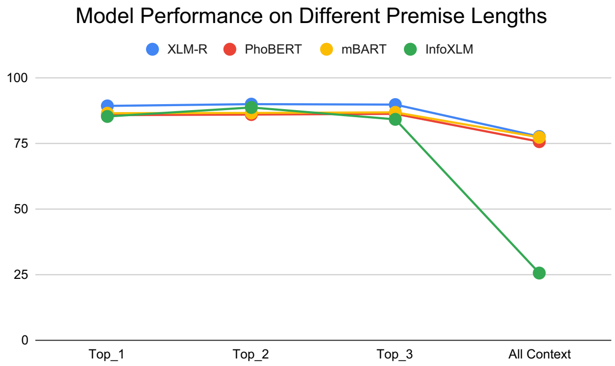


Fig. 8 Performance on different premise lengths

77.57% (All Context). The most visible model InfoXLM decreases from 85.21% to 25.59%. Consequently, the results demonstrate that the longer premise is integral in achieving better performance on the NLI task, but how valuable the information in the premise could affect the performance of models. A similar conclusion is pointed out in the results of the models on the new test.

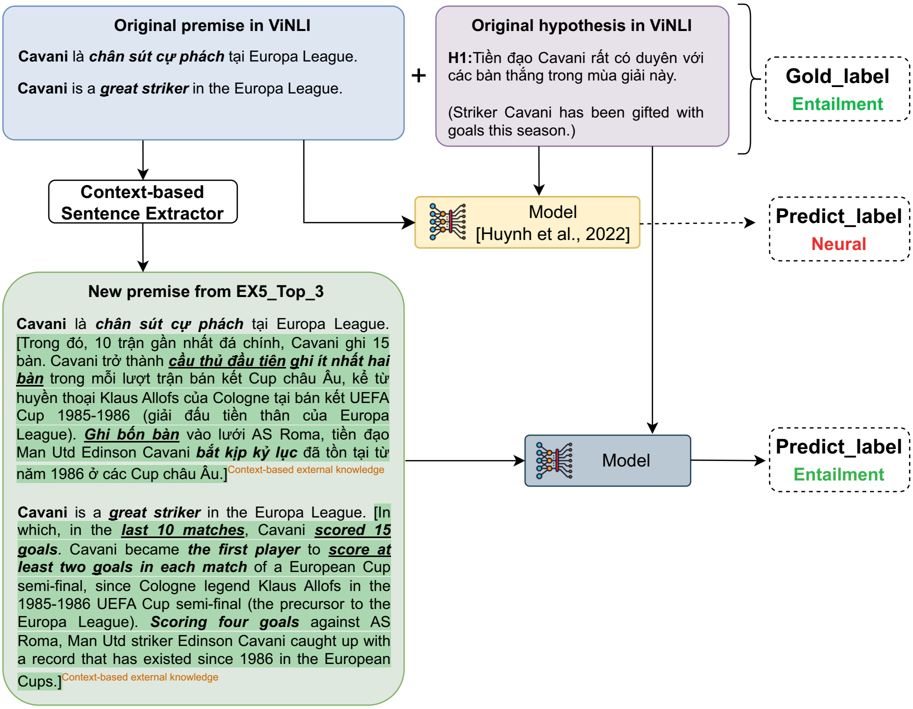
4.3.2 Model performance on different labels

We compare model performance across different labels in [5] and ours. Noteworthy, in the original work with four labels, models such as XLM-R are good at performing on Contradiction and Neutral, but struggling when deciding the relationship of Entrailment. However, as shown in Table 5, our approach can significantly improve the decision of models in examining the Entailment. In particular, the accuracy-based performances of InfoXLM, PhoBERT, mBART, XLM-R, on Entailment increase from 86.33% to 91.21%, 87.96%, 89.31%, 91.47% respectively. Furthermore, our approach not only enhances the performance on the Entailment label, but context-based external knowledge also improves other labels. Specifically, the accuracy on Contradiction, Neutral, and Other labels of XLM-R increased by 2.88%, 1.46%, and 0.41%, respectively. The performance of PhoBERT on Other label enhances 0.43%. mBART improves the Contradiction of 0.4%.

Additionally, Fig. 9 shows one of the prominent cases that our approach can tackle, but the original work did not. Figure 9 demonstrates an example of the challenges brought by analytical reasoning. Specifically, the original premise sentence concerns how good Cavani was in that season and the hypothesis describes the gifted of Cavani in that season. Models

Table 5 Model performance per label in ViNLI

Label	Huynh et al., 2022	Experiment5_Top_1			
		InfoXLM	PhoBERT	mBART	XLM-R
Entailment	86.33	**91.21**	**87.96**	**89.31**	**91.47**
Contradiction	82.98	80.11	78.92	**83.38**	**85.86**
Neutral	80.45	73.01	78.32	75.93	**81.91**
Other	97.34	96.68	**97.75**	97.21	**97.75**


Original premise in ViNLI
+
Original hypothesis in ViNLI
Gold_label
Entailment
Context-based
Sentence Extractor
Model
[Huynh et al., 2022]
Predict_label
Neural
New premise from EX5_Top_3
Model
Predict_label
Entailment
Context-based external knowledge
Context-based external knowledge

Cavani là ***chân sút cự phách*** tại Europa League.

Cavani is a ***great striker*** in the Europa League.

H1: Tiền đạo Cavani rất có duyên với các bàn thắng trong mùa giải này.

(Striker Cavani has been gifted with goals this season.)

Cavani là ***chân sút cự phách*** tại Europa League. [Trong đó, 10 trận gần nhất đá chính, Cavani ghi 15 bàn. Cavani trở thành ***cầu thủ đầu tiên ghi ít nhất hai bàn*** trong mỗi lượt trận bán kết Cup châu Âu, kể từ huyền thoại Klaus Allofs của Cologne tại bán kết UEFA Cup 1985-1986 (giải đấu tiền thân của Europa League). ***Ghi bốn bàn*** vào lưới AS Roma, tiền đạo Man Utd Edinson Cavani ***bắt kịp kỷ lục*** đã tồn tại từ năm 1986 ở các Cup châu Âu.]Context-based external knowledge

Cavani is a ***great striker*** in the Europa League. [In which, in the ***last 10 matches***, Cavani ***scored 15 goals***. Cavani became ***the first player*** to ***score at least two goals in each match*** of a European Cup semi-final, since Cologne legend Klaus Allofs in the 1985-1986 UEFA Cup semi-final (the precursor to the Europa League). ***Scoring four goals*** against AS Roma, Man Utd striker Edinson Cavani caught up with a record that has existed since 1986 in the European Cups.]Context-based external knowledge

Fig. 9 Example of cases that context-based external knowledge can address, which the original did not. The green indicates the original prediction. The red indicates the correct label and our model's prediction. Reasoning clues are highlighted in the context

need to determine the facts after analyzing and deducting. Besides, the lexical overlap between the premise and hypothesis is low. The best model in the original work [5] XLM-R incorrectly chose the Neutral label, while our approach which adds context-based external knowledge mentioning the number of goals and the related records Cavani scored in that season can predict precisely the Entailment label.

4.3.3 Model performance on different inputs

To perform well on the NLI task, humans need more information about the context of the premise and hypothesis, and so do pre-trained language models. Therefore, we experimented with various types of model input, as displayed in the Fig. 10.

Besides our main experiment conducted on context-based external knowledge (i.e. Experiment 5), we designed 4 other different experiments, as described in Section 4.1. Despite the length, and the information that all context provides in Experiment 1, these models did not perform well on the NLI task, which gives 77.26%, 73.72%, 76.53%, 25.59% corresponding to XLM-R, PhoBERT, mBART, InfoXLM. In Experiment 2, Experiment 3, and Experiment 4, after defining how the premise was created in [5], we applied the best IR SXLM-R [18] to get the premise's contextual information. However, compared to Experiment 5 (i.e. our main contribution), these experiments perform worse. In particular, InfoXLM gives the best

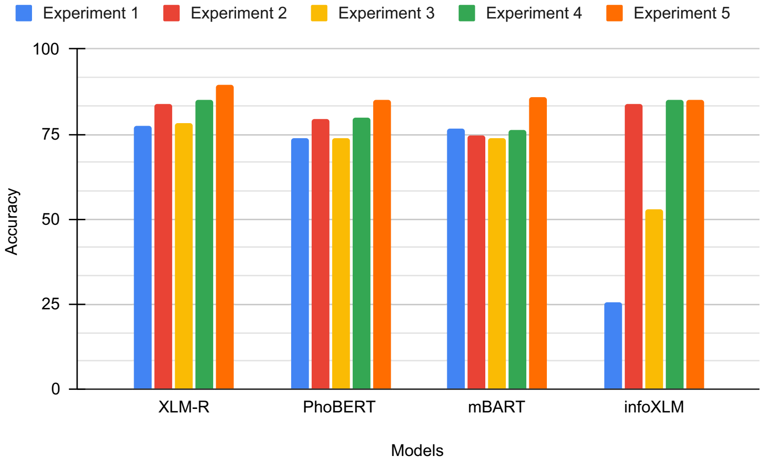


Fig. 10 Model performance on different inputs

accuracy of 83.98%, 85.08% in Experiment 2, and Experiment 4; For Experiment 3, XLM-R outperforms others with 78.37% of accuracy. Thus, we conclude that a longer premise is an indispensable factor in improving model performance in the NLI task; however, the information in the premise should be paid more attention.

5 Conclusion and future works

In this paper, we leverage the only open-domain and high-quality dataset for Vietnamese (ViNLI) to automatically create a long-premise Vietnamese NLI dataset to assess the efficiency of a longer premise. We demonstrate that our approach can obtain better performance in inferring semantic information due to infusing context-based external knowledge created by combining our rules with information retrieval techniques. Therefore, not only do we show that the longer premise is integral in achieving better performance on the NLI task, but we also further indicate how valuable information in the premise could affect the performance of models. Besides, we experiment with both the encoder and encoder-decoder models and point out that for the task, the encoder is more suitable than the transformer. Moreover, our approach successfully achieves state-of-the-art performance for the task of natural language inference with 4 classes on the ViNLI dataset.

However, there are still limitations in our work. In particular, our approach only focuses on the ‘Entailment’ class to improve the performance of the models. Therefore, in the future, designing a more general framework will be paid more attention to exploit relevant knowledge not only for the ‘Entailment’ but also for others based on the given dataset and context. Another direction worth mentioning involves exploring new ways to extract relevant knowledge efficiently to improve performance on the Vietnamese NLI task. Ultimately, we aspire to apply our system to address downstream NLP tasks such as question answering and summarization.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Katz JJ (1972) Semantic Theory. Harper and Row, New York
2. Van Benthem J (2008) A brief history of natural logic
3. Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 632–642. https://doi.org/10.18653/v1/D15-1075 . https://aclanthology.org/D15-1075. Association for Computational Linguistics, Lisbon, Portugal
4. Williams A, Nangia N, Bowman S (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), pp 1112–1122. https://doi.org/10.18653/v1/N18-1101 https://aclanthology.org/N18-1101. Association for Computational Linguistics, New Orleans, Louisiana
5. Huynh TV, Nguyen KV, Nguyen NL-T (2022) ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 3858–3872. https://aclanthology.org/2022.coling-1.339. International Committee on Computational Linguistics, Gyeongju, Republic of Korea
6. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V (2019) Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116
7. Chi Z, Dong L, Wei F, Yang N, Singhal S, Wang W, Song X, Mao X-L, Huang H, Zhou M (2020) Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. arXiv:2007.07834
8. Nguyen DQ, Nguyen AT (2020) Phobert: Pre-trained language models for vietnamese. arXiv:2003.00744
9. Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L (2020) Multilingual denoising pre-training for neural machine translation. Trans Assoc Comput Linguist 8:726–742
10. Mishra A, Patel D, Vijayakumar A, Li XL, Kapanipathi P, Talamadupula K (2021) Looking beyond sentence-level natural language inference for question answering and text summarization. In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1322–1336. https://doi.org/10.18653/v1/2021.naacl-main.104 https://aclanthology.org/2021.naacl-main.104. Association for Computational Linguistics, Online
11. Giunchiglia F (1993) Contextual reasoning. Epistemologia, special issue on I Linguaggi e le Macchine 16:345–364
12. Lai G, Xie Q, Liu H, Yang Y, Hovy E (2017) Race: Large-scale reading comprehension dataset from examinations. arXiv:1704.04683
13. Liu H, Cui L, Liu J, Zhang Y (2021) Natural language inference in context-investigating contextual reasoning over long texts. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 13388–13396
14. Anh HT, Huyen NTM, Lien N et al (2022) Vlsp 2021-vnnli challenge: Vietnamese and english-vietnamese textual entailment. VNU J Sci Comput Sci Commun Eng 38(2)
15. Robertson S, Zaragoza H et al (2009) The probabilistic relevance framework: Bm25 and beyond. Found Trends® Inf Ret 3(4):333–389
16. Paik JH (2013) A novel tf-idf weighting scheme for effective ranking. In: Proceedings of the 36th International ACM SIGIR conference on research and development in information retrieval, pp 343–352
17. Reimers N, Gurevych I (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv:1908.10084
18. Nguyen NT-H, Ha PP-D, Nguyen LT, Van Nguyen K, Nguyen NL-T (2022) Spbertqa: A two-stage question answering system based on sentence transformers for medical texts. In: Knowledge science, engineering and management: 15th international conference, KSEM 2022, Singapore, August 6–8, 2022, Proceedings, Part II, pp 371–382. Springer
19. Dagan I, Glickman O, Magnini B (2006) The pascal recognising textual entailment challenge. In: Machine learning challenges workshop, pp 177–190. Springer
20. Giampiccolo D, Magnini B, Dagan I, Dolan WB (2007) The third pascal recognizing textual entailment challenge. In: Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing, pp 1–9

21. Bentivogli L, Clark P, Dagan I, Giampiccolo D (2009) The fifth pascal recognizing textual entailment challenge. In: TAC
22. Bentivogli L, Clark P, Dagan I, Giampiccolo D (2010) The sixth pascal recognizing textual entailment challenge. In: TAC
23. Bentivogli L, Clark P, Dagan I, Giampiccolo D (2011) The seventh pascal recognizing textual entailment challenge. In: TAC
24. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv:1804.07461
25. Khot T, Sabharwal A, Clark P (2018) Scitail: A textual entailment dataset from science question answering. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
26. Welleck S, Weston J, Szlam A, Cho K (2018) Dialogue natural language inference. arXiv:1811.00671
27. Hu H, Richardson K, Xu L, Li L, Kübler S, Moss LS (2020) Ocnli: Original chinese natural language inference. arXiv:2010.05444
28. Mahendra R, Aji AF, Louvan S, Rahman F, Vania C (2021) Indonli: A natural language inference dataset for indonesian. arXiv:2110.14566
29. Wijnholds G, Moortgat M (2021) Sicknl: A dataset for dutch natural language inference. arXiv:2101.05716
30. Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist 2:67–78
31. Lai A, Bisk Y, Hockenmaier J (2017) Natural language inference from multiple premises. arXiv:1710.02925
32. Nie Y, Williams A, Dinan E, Bansal M, Weston J, Kiela D (2019) Adversarial nli: A new benchmark for natural language understanding. arXiv:1910.14599
33. Elman JL (1990) Finding structure in time. Cogn Sci 14(2):179–211
34. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
35. Chen Q, Zhu X, Ling Z, Wei S, Jiang H, Inkpen D (2016) Enhanced lstm for natural language inference. arXiv:1609.06038
36. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
38. Devlin J, Chang M-W, Lee K, Toutanova K (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), pp 4171–4186. https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423. Association for Computational Linguistics, Minneapolis, Minnesota
39. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv:1910.13461
40. Roberts A, Raffel C (2020) Exploring transfer learning with t5: the text-to-text transfer transformer. Accessed on, 23–July
41. Phan L, Tran H, Nguyen H, Trinh TH (2022) Vit5: Pretrained text-to-text transformer for vietnamese language generation. arXiv:2205.06457
42. Vu HX, Van Tai N, Khoa PTK, Van Thin D, Hao DN, Ngan NLT (2022) vnnli-vlsp 2021: Vietnamese and english-vietnamese textual entailment based on pre-trained multilingual language models. VNU J Sci Comput Sci Commun Eng 38(2)
43. Luan ND, Kien NLH, Van Thin D, Hao DN, Ngan NLT (2022) vnnli-vlsp2021: An empirical study on vietnamese-english natural language inference based on pretrained language models with data augmentation. VNU J Sci Comput Sci Commun Eng 38(2)
44. Nguyen CT, Nguyen DT (2022) Building a vietnamese dataset for natural language inference models. SN Comput Sci 3(5):395
45. Duong Q-L, Nguyen D-V, Nguyen NL-T (2022) Leveraging semantic representations combined with contextual word representations for recognizing textual entailment in vietnamese. In: 2022 9th NAFOSTED conference on information and computer science (NICS), pp 47–52. IEEE
46. Yang M, Tu W, Qu Q, Zhou W, Liu Q, Zhu J (2019) Advanced community question answering by leveraging external knowledge and multi-task learning. Knowl-Based Syst 171:106–119
47. Wang X, Kapanipathi P, Musa R, Yu M, Talamadupula K, Abdelaziz I, Chang M, Fokoue A, Makni B, Mattei N, et al (2019) Improving natural language inference using external knowledge in the science questions domain. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 7208–7215

48. Deng Y, Xie Y, Li Y, Yang M, Lam W, Shen Y (2021) Contextualized knowledge-aware attentive neural network: Enhancing answer selection with knowledge. ACM Trans Inf Syst (TOIS) 40(1):1–33
49. Tahir M, Halim Z, Waqas M, Tu S (2023) On the effect of emotion identification from limited translated text samples using computational intelligence. Int J Comput Intell Syst 16(1):107
50. Halim Z, Waqar M, Tahir M (2020) A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. Knowl-Based Syst 208:106443
51. Tahir M, Halim Z, Rahman AU, Waqas M, Tu S, Chen S, Han Z (2022) Non-acted text and keystrokes database and learning methods to recognize emotions. ACM Trans Multimed Comput Commun Appl (TOMM) 18(2):1–24
52. Chen Q, Zhu X, Ling Z-H, Inkpen D, Wei S (2017) Neural natural language inference models enhanced with external knowledge. arXiv:1711.04289
53. Kapanipathi P, Thost V, Patel SS, Whitehead S, Abdelaziz I, Balakrishnan A, Chang M, Fadnis K, Gunasekara C, Makni B., et al (2020) Infusing knowledge into the textual entailment task using graph convolutional networks. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 8074–8081
54. Peng C, Xia F, Naseriparsa M, Osborne F (2023) Knowledge graphs: Opportunities and challenges. Artif Intell Rev 1–32
55. Chen D, Fisch A, Weston J, Bordes A (2017) Reading wikipedia to answer open-domain questions. arXiv:1704.00051
56. Karpukhin V, Oğuz B, Min S, Lewis P, Wu L, Edunov S, Chen D, Yih W-t (2020) Dense passage retrieval for open-domain question answering. arXiv:2004.04906
57. Wang S, Yu M, Guo X, Wang Z, Klinger T, Zhang W, Chang S, Tesauro G, Zhou B, Jiang J (2018) R 3: Reinforced ranker-reader for open-domain question answering. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
58. Lee J, Seo M, Hajishirzi H, Kang J (2019) Contextualized sparse representations for real-time open-domain question answering. arXiv:1911.02896
59. Hildebrand AS, Eck M, Vogel S, Waibel A (2005) Adaptation of the translation model for statistical machine translation based on information retrieval. In: Proceedings of the 10th EAMT conference: practical applications of machine translation
60. Nair S, Yang E, Lawrie D, Mayfield J, Oard DW (2022) Learning a sparse representation model for neural clir. Design of Experimental Search and Information REtrieval Systems (DESIRES)
61. Kim M-Y, Rabelo J, Okeke K, Goebel R (2022) Legal information retrieval and entailment based on bm25, transformer and semantic thesaurus methods. Rev Socionetwork Str 16(1):157–174
62. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
63. Henderson M, Al-Rfou R, Strope B, Sung Y-H, Lukács L, Guo R, Kumar S, Miklos B, Kurzweil R (2017) Efficient natural language response suggestion for smart reply.arXiv:1705.00652
64. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692
65. Vu T, Nguyen DQ, Nguyen DQ, Dras M, Johnson M (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, pp 56–60. https://doi.org/10.18653/v1/N18-5012 https://aclanthology.org/N18-5012. Association for Computational Linguistics, New Orleans, Louisiana
66. Loshchilov I, Hutter F (2017) Decoupled Weight Decay Regularization. arXiv:1711.05101

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.