VNU Journal of Science:
Computer Science and Communication Engineering

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

Original Article

ViMRC VLSP 2021: XLM-R Versus PhoBERT on Vietnamese Machine Reading Comprehension

Nguyen Duy Nhat*, Do Nguyen Thuan Phong

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

Received 28 December 2021
Revised 26 April 2022; Accepted 5 May 2022

Abstract: The development of industry 4.0 in the world is creating challenges in Artificial Intelligence (AI) in general and Natural Language Processing (NLP) in particular. Machine Reading Comprehension (MRC) is an NLP task with real-world applications that require machines to determine the correct answers to questions based on a given document. MRC systems must not only answer questions when possible but also determine when the document supports no answer and abstain from answering. In this paper, we present our proposed system to solve this task at the VLSP shared task 2021: Vietnamese Machine Reading Comprehension with UIT-ViQuAD 2.0. We present the MRC4MRC model to address that task. The model is made up of two separate components with the same automatic reading function in the MRC4MRC model. Our MRC4MRC based on the XLM-RoBERTa pre-trained language model achieves 79.13% in F1-score (F1) and 69.72% in EM (Exact Match) on the public test set. Our experiments also show that the XLM-RoBERTa language model is better than the powerful PhoBERT language model on UIT-ViQuAD 2.0.

Keywords: Machine Reading Comprehension, Vietnamese, XLM-Roberta.

1. Introduction

The past decade has seen tremendous development in MRC, including an increase in corpus numbers across languages and significant progress in techniques. Indeed, datasets and models impact the performance of the MRC tasks. For the datasets, researchers have developed many MRC datasets such as SQuAD 2.0 [1], CMRC2019 [2], FQuAD [3], UIT-ViNewsQA [4], UITViWikiQA [5], UIT-ViQuAD [6], and MLQA [7]. Besides developing MRC datasets, various significant neural network-based methods have been proposed and made significant advances in the MRC problem, such as FastQA [8], QANet [9], and BERT [10].

* Corresponding author.
E-mail address: 18520118@gm.uit.edu.vn

https://doi.org/10.25073/25881086/vnucsce.334

To develop and improve the MRC task in Vietnamese, VLSP Shared Task 2021 opens with Vietnamese Machine Reading Comprehension. In this task, we must help the MRC model correctly predict the question's answer based on the document and predict unanswerable questions. Table 1 shows several examples for this task.

In this paper, we have two primary contributions described as follows:

- We propose a new combined MRC model (MRC4MRC) inspired by solving the Vietnamese MRC task held at VLSP Shared Task 2021. Our proposed model achieves superior performance than PhoBERT and XLM-Roberta models and scored 6th on the private-test set.
- We analyze the answerable and unanswered questions to show two conclusions for the Vietnamese MRC task: the XLM-Roberta model is better than the PhoBERT model, and the performance of the MRC model (both XLMRoberta and PhoBERT) is improved by using BiLSTM as the final layer instead of the linear layer.

The rest of the paper is organized as follows. In Section 2, we presented the related work. In Section 3, we describe the data pre-processing process. In Section 4, we described our MRC4MRC model. In Section 5, we presented the experimental results. Finally, the last section gives conclusions about the work.

Table 1. An example of answerable and unanswerable questions in VLSP Shared Task 2021 Vietnamese Machine Reading Comprehension

Context: Malaysia có nguồn gốc từ các vương quốc Mã Lai hiện diện trong khu vực, và từ thế kỷ XVIII, các vương quốc này bắt đầu lệ thuộc vào Đế quốc Anh. Các lãnh thổ đầu tiên của Anh Quốc được gọi là Các khu định cư Eo biển. Các lãnh thổ tại Malaysia bán đảo được hợp nhất thành Liên hiệp Malaya vào năm 1946. Malaya được tái cấu trúc thành Liên bang Malaya vào năm 1948, và giành được độc lập vào ngày 31 tháng 8 năm 1957. Malaya hợp nhất với Bắc Borneo, Sarawak, và Singapore vào ngày 16 tháng 9 năm 1963,với từ si được thêm vào quốc hiệu mới là Malaysia. Đến năm 1965, Singapore bị trục xuất khỏi liên bang. (*Malaysia is descended from the Malay kingdoms present in the region, and from the eighteenth century, these kingdoms became subordinate to the British Empire. The first British territories were known as the Straits Settlements. Territories in Peninsular Malaysia were incorporated into the Union of Malaya in 1946.* *Malaya was restructured into the Federation of Malaya in 1948 and gained independence on 31 August1957. Malaya merged with North Borneo, Sarawak, and Singapore on 16 September 1963, with the word sibeing added to the new national title, Malaysia. In 1965, Singapore was expelled from the federation*).
Question 1 (**Answerable Question**): Các lãnh thổ đầu tiên của Anh Quốc tại Malaysia được gọi là gì? (*What were the first British territories in Malaysia called?*). **Answer**: **được gọi là Các khu định cư Eo biển** (*were known as the Straits Settlements*).
Question 2 (**Unanswerable Question**): Các lãnh thổ kế tiếp của Anh Quốc tại Malaysia được gọi là gì? (*What are the next British territories in Malaysia called?*). Answer: ""

2. Related Work

The development of MRC problem and Question Answering (QA) systems require the integration of datasets and models. In this section, we briefly review the datasets and models commonly used in the MRC problem and the QA system.

2.1. Machine Reading Comprehension and Question Answering Datasets

There are many datasets about MRC and QA in different languages in the world.

MLQA [7] is a benchmark dataset for evaluating crosslingual question answering performance (English, Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese). MKQA [11] is an open-domain question answering evaluation set comprising 10k question-answer pairs aligned across 26 typologically diverse languages. XQA [12] is a data which consists of a total amount of 90k question-answer pairs in nine languages for cross-lingual open-domain question answering. In English, Rajpurkar et al. published the dataset named SQuAD 2.0 [1]. SQuAD 2.0 combines more than 50,000 unanswerable questions with more than 100,000 answerable questions from the existing SQuAD 1.1 dataset [13] to create a large dataset (more than 150,000 questions) widely used in MRC tasks and QA systems. Yang et al. published the WiKiQA dataset [14] that was collected and annotated for research on open-domain question answering. Joshi et al. presented the TriviaQA dataset [15], a challenging reading comprehension dataset containing over 650K question-answer-evidence triples. CMRC 2019 [2] is a Chinese Machine Reading Comprehension. Specifically, CMRC 2019 is a sentence cloze-style machine reading comprehension dataset that aims to evaluate the sentence-level inference ability. FQuAD [3] is a French Native Reading Comprehension dataset of questions and answers on a set of Wikipedia articles that consists of 25,000+ samples.

In Vietnamese, MRC datasets have also been developed in recent years. ViNewsQA [4] is a new corpus for the Vietnamese language to evaluate healthcare reading comprehension models consisting of 22,057 question-answer pairs. UIT-ViWikiQA [5] is the first dataset for evaluating sentence extraction-based machine reading comprehension in the Vietnamese language. UITViQuAD [6] is the first extraction-based MRC dataset that comprises over 23,000 human-generated question-answer pairs based on 5,109 passages of 174 Vietnamese articles from Wikipedia. UIT-ViCoQA [16] is a new corpus for conversational reading in Vietnamese that includes 10,000 questions with answers to more than 2,000 conversations about health news articles. ViMMRC [17] is a dataset which consists of 2,783 pairs of multiple-choice questions and answers based on 417 Vietnamese texts which are commonly used for teaching reading comprehension for elementary school pupils.

2.2. Machine Reading Comprehension Models

In the process of researching and solving the MRC task, many models were created, such as QANet [9], FusionNet [18], DrQA [19], FastQA [8]. Most recently, BERT [10] and XLM-Roberta [20], which are potent models trained on multiple languages, have obtained state-of-the-art performances on MRC problems ([5, 21]). Besides, a variant of BERT called PhoBERT [22] is a monolingual pre-trained language model for Vietnamese. MRC models are combined with document retrievers to create QA systems in Vietnamese and other languages such as ViQAS [23], XLMRserini [24], DrQA [19], and BERTserini [25].

3. Pre-Processing

Before training the model, we pre-process the data based on the rule approach with the question word conversion rules [23]. We change the word to ask in the questions based on five algorithms corresponding to the five types of questions, including When questions, Where questions, Who questions, Why questions, and What questions. Questions of the same type with various question words make the model less efficient [5], so we use a set of rules to reduce the variety of question words.

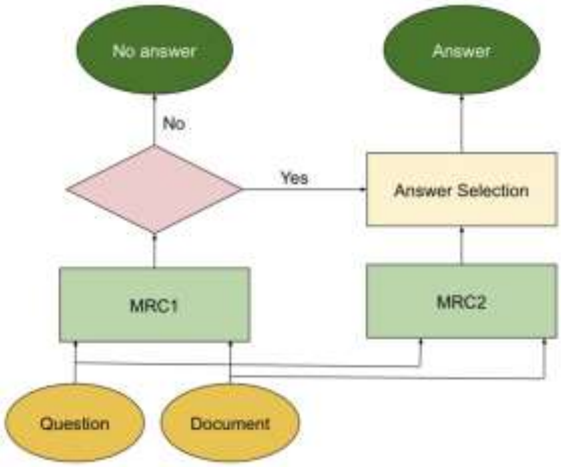


Figure 1. Architecture of MRC4MRC model for machine reading comprehension task.

4. MRC4MRC Model for Vietnamese Machine Reading Comprehension

In this paper, we propose a model MRC4MRC based on a combination of two answer extractor components of the same type (called MRC1 and MRC2). Given a document and a preprocessed question, two components MRC1 and MRC2, perform unanswerable question detection and provide answers to answerable questions. MRC1 and MRC2 are two separate components with the same automatic reading function in the MRC4MRC model, and these two components have the same architecture and only different model training parameters. The overview of the MRC4MRC model is depicted in Figure 1. These components are described below.

4.1. Machine Reading Comprehension Components

Inspired by the superiority of the XLM-R model on multiple tasks [20, 23], we implement two components of answer extraction (MRC1 and MRC2) based on the transfer learning approach. The XLM-R model was trained on previous multilingual data and enhanced to solve for answer extraction with the ViQuAD 2.0 dataset [26] provided by the VLSP competition. To feed the task of extracting the answers into XLM-R, we put the question and the document into the input. At the input, we add two unique tokens [CLS] and [SEP]. [CLS] is used for classification purposes to classify answerable and unanswerable questions in the MRC1 component. [SEP] is a unique token that separates two pieces of text input. It is located in the position after the question.

Through the XLM-Roberta layer, the hidden vector representation Ci is computed for each token i. Inspired by Nguyen et al. [23], we add the Bidirectional LSTM layer at the output of the XLM-Roberta layer. For each token i, we process forward the hidden state $\overrightarrow{C_i}$ and backward the hidden state $\overleftarrow{C_i}$ to obtain Li and Li. We set the hidden dimension to 2 because each token contains two values as the probability of starting and ending the answer. This way is widely used in modern MRC models, with the last layer being both Linear and Bi-LSTM. At each token i, we connect them to obtain the final state Li..

$$\overrightarrow{L_i} = LSTM\left(\overrightarrow{C_i}\right) \qquad (1)$$

$$\overleftarrow{L_i} = LSTM\left(\overleftarrow{C_i}\right) \qquad (2)$$

$$L_i = \overrightarrow{L_i} + \overleftarrow{L_i} \qquad (3)$$

At each token i, we get the start of the answer (S) as the first element of the final state Li and the end of the answer (E) as the second element of the final state Li. Start and end probabilities at each token i are calculated via softmax function as follows.

$$Pstart_i = \frac{\exp\,(S.T_i)}{\sum_j (S.T_j)} \qquad (4)$$

$$Pend_i = \frac{\exp\,(E.T_i)}{\sum_j (E.T_j)} \qquad (5)$$

The training process is done through the optimization of the goal below.

$$\sum_k \left(\log(Pstart_{sk}) + \log(Pend_{ek})\right) \qquad (6)$$

Where *sk* and *ek* are the indices of the exact start and end index of the sample *k*.

In inference time, the predicted answer is found by finding the candidates for the starting and ending positions by optimizing the following problem.

$$\max_{i,\,j}\left(S.T_i + E.T_j\right) \qquad (7)$$

On the MRC1 model, the model needs to recognize unanswerable questions. The token [CLS] is included in the model as the first token. A null answer score is calculated by formula 8, where C is the final hidden state corresponding to the token [CLS].

$$S.C + E.C \qquad (8)$$

Finally, the MRC1 model predicts the unanswerable question in case the following condition is satisfied.

$$S.C + E.C \geq \max_{i,\,j}\left(S.T_i + E.T_j\right) \qquad (9)$$

4.2. Answer Selection Component

The model input includes two answers from the MRC1 and MRC2 models along with their respective scores. The score of the answers is calculated from the machine reading comprehension components through formula number 7. In case the MRC1 model predicts a answerable question, the component that chooses the answer is activated. Otherwise, this component is not working. The output of this component is the answer with the higher score.

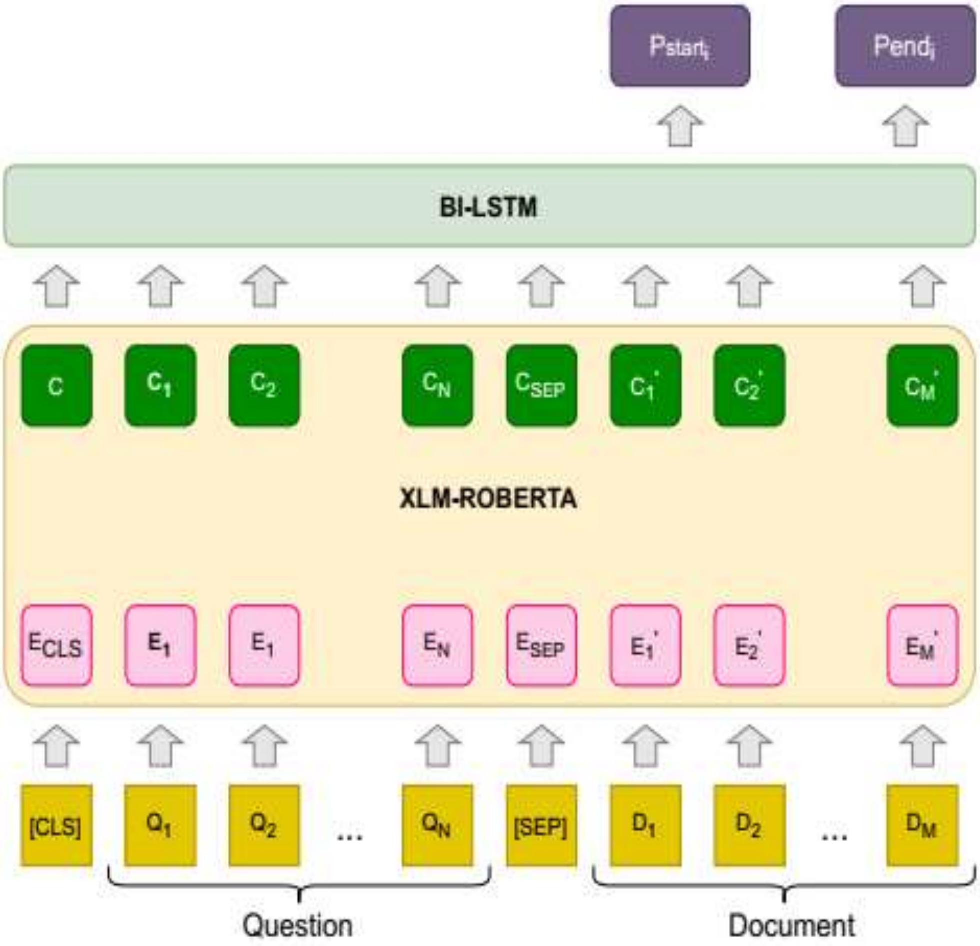


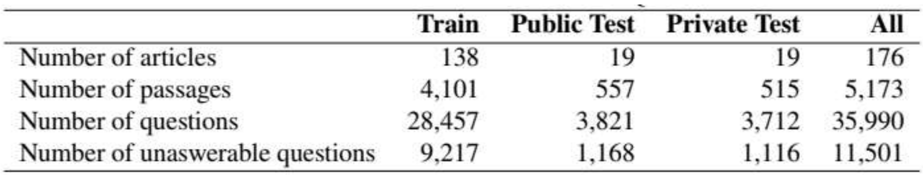
Figure 2. The answer-extractor component of the MRC4MRC model is built based on the XLM-Roberta model.

5. Experiments

5.1. Dataset

We use the UIT-ViQuAD 2.0 dataset [26] provided by VLSP Share Task 2021. The training set includes 28457 questions in JSON format, a public-test set of 3821 questions, and a privatetest set of 3712 questions. The dataset articles are taken from the Wikipedia website, and humans annotate the questions and answers in Vietnamese. The detailed information about the dataset is described in Table 2. The dataset has unanswerable questions and answerable questions. In particular, unanswerable questions are approximately 32% of the total number of questions in the dataset. Some data examples are shown in Table 1.

Table 2. The detailed information about the UIT-ViQuAD 2.0



	Train	Public Test	Private Test	All
Number of articles	138	19	19	176
Number of passages	4,101	557	515	5,173
Number of questions	28,457	3,821	3,712	35,990
Number of unaswerable questions	9,217	1,168	1,116	11,501

5.2. Experimental Settings

We use a single NVIDIA Tesla K80 via Google Colaboratory to train our model. The baseline configuration provided by HuggingFace1 is used to fine-tune the machine reading comprehension components of the MRC4MRC model and other models implemented by us for comparison such as: PhoBERT$_{large}$ (with Linear Layer or BiLSTM layer is the final layer), and XLM-Roberta$_{large}$ (with Linear Layer or BiLSTM layer is the final layer). We set $epochs = 2$, $learning - rate = 2e^{-5}$, and a maximum string length of 384. Besides, we also implement other models to compare the performance of the models with our proposed model. Additional models we use include: the XLM-Roberta$_{large}$ model [20], and the pre-trained language model for Vietnamese PhoBERT$_{large}$ model [22]. Both implementation models for comparison work are developed with two different final layers, including Linear and BiLSTM. Before training the MRC model using PhoBERT as a pre-trained model, we follow the request of the author of PhoBERT to conduct word segmentation with the VnCoreNLP toolkit [27].

We conduct training of the MRC4MRC model in two separate and independent phases, including MRC1 and MRC2. We train MRC1 with all the data in the training set of the UITViQuAD 2.0 dataset and MRC2 with the answerable questions set of the UIT-ViQuAD 2.0 dataset. The two models are linked together through the answer selection component we described in Section 4.

5.3. Evaluation Metrics

The two evaluations metrics used by the MRC task organizer to evaluate the performance of the model are F1-score (F1) and EM. Considering two answers, including the gold and predicted answers, if the two answers are the same, EM is set to 1, and EM is set to 0 otherwise.

$$EM = \frac{Number \text{ of correctly predicted questions}}{Number \text{ of questions}} \quad (10)$$

F1 measures the overlap between the gold answer and the predicted answer, calculated by Formula 11, where Precision is the ratio of the number of correctly predicted tokens to the total number of predicted tokens, and Recall is the ratio of the number of correctly predicted tokens to the total number of tokens of the gold answer. F1 is used as the key evaluation metric to rank the result of teams.

$$F_1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

5.4. Experimental Results

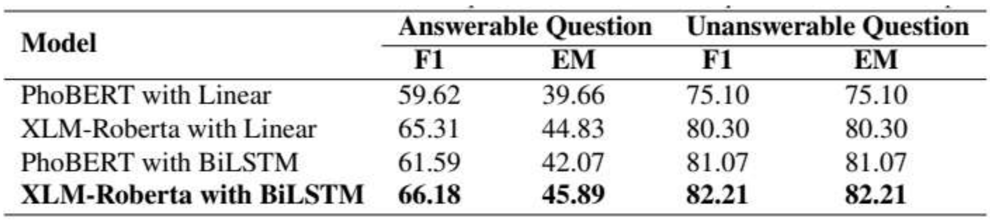
5.4.1. Model Performance on the Set of Answerable and UnanSwerable Questions

To select two machine reading components in the MRC4MRC model, we deploy several models to choose the one with the best ability to recognize unanswered questions and the best one in extracting answers to the answerable questions are on the development set. The development set is divided by ourselves with more than 2000 questions to compare the performance of the model. With the results shown in Table 4, the models using the last layer of BiLSTM have better performance than the models using the final layer of Linear and are on average 2.68% higher for F1 and 2.84% for EM. The XLM-Roberta$_{large}$ model achieved outstanding performance, especially XLM-Roberta$_{large}$ has the last layer being BiLSTM achieving the highest performance on both the set of answerable and unanswerable questions. For the unanswerable questions, the XLM-Roberta$_{large}$+BiLSTM model achieve a recognition efficiency of 82.21%. On the answerable questions, the XLM-Roberta$_{large}$+BiLSTM model also achieve the best performance with F1 of 66.18% and EM of 45.89%. Therefore, we choose the XLM-Roberta$_{large}$+BiSLTM model for both MRC1 and MRC2 components of our proposed MRC4MRC system to provide the best overall system performance.

5.4.2. Performance of Models

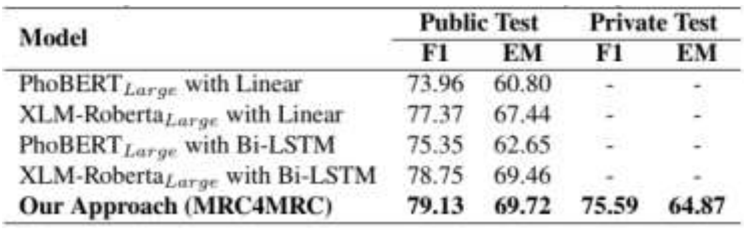
In this section, we show the task results in Table 3. The models are implemented based on the XLM-Roberta$_{large}$ pretrained multilingual model, outperforming models implemented based on PhoBERT$_{large}$ pre-trained monolingual model. Our proposed model (MRC4MRC) achieves the best performance compared to other models to solve the Vietnamese MRC task. In detail, the MRC4MRC model performed an F1-score of 79.13% and an EM of 69.72% on the public test set. Our model (MRC4MRC) versus the original XLM-R model with 1.16% higher F1-score and 2.28% higher EM on the public test set. Adding the Bi-LSTM layer also helps our model increase 1.38% F1-score and 2.02% EM on the public test set of the UIT-ViQuAD dataset.

Table 3. Performance of the models in the unanswerable question and the answerable question set in the development set



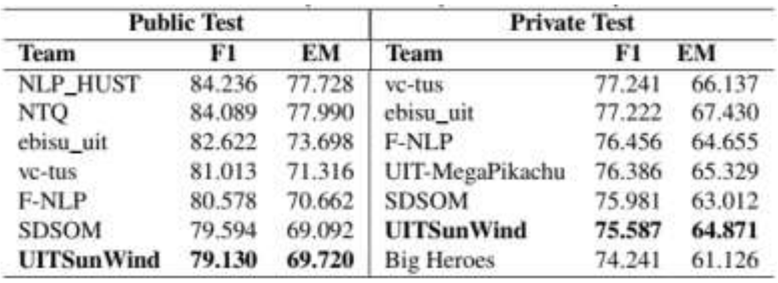
Model	Answerable Question		Unanswerable Question	
	F1	EM	F1	EM
PhoBERT with Linear	59.62	39.66	75.10	75.10
XLM-Roberta with Linear	65.31	44.83	80.30	80.30
PhoBERT with BiLSTM	61.59	42.07	81.07	81.07
XLM-Roberta with BiLSTM	**66.18**	**45.89**	**82.21**	**82.21**

Table 4. Our experimental results on the Vietnamese Machine Reading Comprehension task



Model	Public Test		Private Test	
	F1	EM	F1	EM
PhoBERT$_{Large}$ with Linear	73.96	60.80	-	-
XLM-Roberta$_{Large}$ with Linear	77.37	67.44	-	-
PhoBERT$_{Large}$ with Bi-LSTM	75.35	62.65	-	-
XLM-Roberta$_{Large}$ with Bi-LSTM	78.75	69.46	-	-
Our Approach (MRC4MRC)	**79.13**	**69.72**	**75.59**	**64.87**

Table 5. Table of results of the top 7 teams on the public test set and the private test set



Public Test			Private Test		
Team	F1	EM	Team	F1	EM
NLP_HUST	84.236	77.728	vc-tus	77.241	66.137
NTQ	84.089	77.990	ebisu_uit	77.222	67.430
ebisu_uit	82.622	73.698	F-NLP	76.456	64.655
vc-tus	81.013	71.316	UIT-MegaPikachu	76.386	65.329
F-NLP	80.578	70.662	SDSOM	75.981	63.012
SDSOM	79.594	69.092	**UITSunWind**	**75.587**	**64.871**
UITSunWind	**79.130**	**69.720**	Big Heroes	74.241	61.126

As shown in Table 5, the MRC4MRC model achieves rank seventh when testing on the public test set. On a private-test set, our proposed model has a sixth performance when using F1 as the metric to rank teams.

6. Conclusion and Future Work

This paper presented our approach to solving the Vietnamese machine reading comprehension task proposed at VLSP ShareTask 2021. We implemented the MRC4MRC model to detect the unanswerable question and extract the answerable questions based on the provided documents. We participated in the Vietnamese MRC task and evaluated the performance of our system on the UIT-ViQuAD 2.0 dataset proposed by MRC task organizers at the VLSP Shared Task 2021. As above, our result is an F1-score of 75.59% and EM of 64.87%, ranking the 7th of the scoreboard on the private-test set. We also show that the performance of the XLM-Roberta model outperforms the PhoBERT model for the Vietnamese MRC tasks. When using BiLSTM as the final layer instead of Linear, the performance of the MRC models (XLM-Roberta and PhoBERT) is improved on both F1 and EM. In future work, we plan to address this task in different ways to enhance performance. We will investigate experiments on a deep neural networks approach, a transfer learning approach, and an ensemble model approach. We also analyze experimental results on this task to select the efficient approach to boost the result of the Vietnamese machine reading comprehension task.

References

[1] P. Rajpurkar, R. Jia, P. Liang, Know What You Don't Know: Unanswerable Questions for Squad, 2018, arXiv:1806.03822.

[2] Y. Cui, T. Liu, W. Che, L. Xiao, Z. Chen, W. Ma, S. Wang, G. Hu, A Span-Extraction Dataset for Chinese Machine Reading Comprehension, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5883-588, https://doi.org/10.18653/ v1/D19-1600.

[3] M. d'Hoffschmidt, W. Belblidia, T. Brendlé, Q. Heinrich, M. Vidal, Fquad: French Question Answering Dataset, 2020, arXiv:2002.06071.

[4] K. V. Nguyen, T. V. Huynh, D. V. Nguyen, A. G. T. Nguyen, N. L. T. Nguyen, New Vietnamese Corpus For Machine Reading Comprehension of Health News Articles, 2021, arXiv:2006.11138.

[5] P. N. T. Do, N. D. Nguyen, T. V. Huynh, K. V. Nguyen, A. G. T. Nguyen, N. L. T. Nguyen, Sentence Extraction-Based Machine Reading Comprehension for Vietnamese, Knowledge Science, Engineering and Management, 2021, pp. 511-523.

[6] K. Nguyen, V. Nguyen, A. Nguyen, N. Nguyen, A Vietnamese Dataset for Evaluating Machine Reading Comprehension, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 2595-2605.

[7] P. Lewis, B. Og˘ uz, R. Rinott, S. Riedel, H. Schwenk, Mlqa: Evaluating Cross-Lingual Extractive Question Answering, arXiv: 1910.07475.

[8] D. Weissenborn, G. Wiese, L. Seiffe, Making Neural Qa as Simple as Possible but not Simpler, arXiv:1703.04816.

[9] A. W. Yu, D. Dohan, M. T. Luong, R. Zhao, K. Chen, M. Norouzi, Q. V. Le, Qanet: Combining Local Convolution with Global Self-Attention for Reading Comprehension, arXiv:1804.09541.

[10] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019, arXiv: 1810.04805.

[11] S. Longpre, Y. Lu, J. Daiber, Mkqa: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering arXiv:2007. 15207.

[12] J. Liu, Y. Lin, Z. Liu, M. Sun, XQA: A Cross-Lingual Open-Domain Question Answering Dataset, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2358-2368.

[13] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ Questions For Machine Comprehension of Text. arXiv:1606.05250.

[14] Y. Yang, W. T. Yih, C. Meek, WikiQA: A Challenge Dataset for Open-Domain Question Answering, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for

Computational Linguistics, Lisbon, Portugal, 2015, pp. 2013-2018.

[15] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, Triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, 2017, arXiv:1705.03551.

[16] S. T. Luu, M. N. Bui, L. D. Nguyen, K. V. Tran, K. V. Nguyen, N. L. T. Nguyen, Conversational Machine Reading Comprehension for Vietnamese Healthcare Texts, Communications in Computer and Information Science, 2021, pp. 546-558.

[17] K. V. Nguyen, K. V. Tran, S. T. Luu, A. G. T. Nguyen, N. L. T. Nguyen, Enhancing Lexical-Based Approach with External Knowledge For Vietnamese Multiple-Choice Machine Reading Comprehension, IEEE Access, Vol. 8, 2020, pp. 201404-201417.

[18] H. Y. Huang, C. Zhu, Y. Shen, W. Chen, Fusionnet: Fusing Via Fully-Aware Attention with Application to Machine Comprehension, 2018, arXiv: 1711.07341.

[19] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading Wikipedia to Answer Open-Domain Questions, 2017, arXiv:1704.00051.

[20] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, UnsuperVised Cross-Lingual Representation Learning At Scale, 2020, arXiv: 1911.02116.

[21] H. Choi, J. Kim, S. Joe, S. Min, Y. Gwon, Analyzing Zero-Shot CrossLingual Transfer in Supervised Nlp Tasks, 2021, arXiv:2101.10649.

[22] D. Q. Nguyen, A. T. Nguyen, Phobert: Pre-Trained Language Models for Vietnamese, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1037-1042.

[23] K. V. Nguyen, P. N. T. Do, N. D. Nguyen, A. G. T. Nguyen, N. L. T. Nguyen, Multi-Stage Transfer Learning with Bertology-Based Language Models for Question Answering System in Vietnamese, VLSP 2021, 2021.

[24] K. V. Nguyen, P. N. T. Do, N. D. Nguyen, A. G. T. Nguyen, T. V. Huynh, N. L. T. Nguyen, Xlmrserini: Open-Domain Question Answering on Vietnamese Wikipedia-Based Textual Knowledge Source, 2021.

[25] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, J. Lin, End-to-end Open-domain Question Answering with Bertserini, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 72-77.

[26] N. V. Kiet, T. Q. Son, N. T. Luan, H. V. Tin, L. T. Son, N. L. T. Ngan, VLSP 2021 ViMRC Challenge: Vietnamese Machine Reading Comprehension, VLSP 2021, 2021.

[27] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, M. Johnson, 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, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 56-60, https://doi.org/10.18653/v1/N18-5012.