F-NLP at VLSP 2021-MRC Shared Task: Joint Learning and Ensemble Method for Vietnamese Machine Reading Comprehension

Phuc Phan Xuan [1,4*], **Tung Nguyen Ky** [2,4†], **Duong Nguyen Hai** [2,4†], **Tri Duong Minh** [3,4†]

[1] FTECH CO., LTD, Ha Noi, Viet Nam
[2] Hanoi University of Science and Technology, Ha Noi, Viet Nam
[3] Danang University of Science and Technology, Da Nang, Viet Nam
[4] {phanxuanphucnd, nguyenkytungcntt04, nhd.a1.chel, minhtridn1999}@gmail.com

Abstract

Machine reading comprehension (MRC) is a challenging Natural Language Processing (NLP) research field and wide real-world applications. The great progress of this field in recents is mainly due to the emergence of few datasets for machine reading comprehension tasks with large sizes and deep learning. For the Vietnamese language, some datasets, such as UIT-ViQuAD (Kiet et al., 2020) and UIT-ViNewsQA (Kiet et al., 2021a), most recently, UIT-ViQuAD 2.0 (Kiet et al., 2021b) - a dataset of the competitive VLSP 2021-MRC Shared Task[1]. MRC systems must not only answer questions when necessary but also tactfully abstain from answering when no answer is available according to the given passage. In this paper, we proposed two types of joint models for answerability prediction and pure-MRC prediction with/ without a dependency mechanism to learn the correlation between a start position and end position in pure-MRC output prediction. Besides, we use ensemble models and a verification strategy by voting the best answer from the top K answers of different models. Our proposed approach is evaluated on the benchmark VLSP 2021-MRC Shared Task challenge dataset UIT-ViQuAD 2.0 (Kiet et al., 2021b) shows that our approach is significantly better than the baseline.

Index Terms: Machine Reading Comprehension, Question Answering, Natural Language Processing, Joint learning, Ensemble models.

1 Introduction

With the rapid development of NLP, natural language understanding (NLU) has aroused broad interests, and a series of NLU tasks have emerged. In order to teach computers to read the text and understand the meaning of the text, researchers have conducted machine reading comprehension (MRC) research. The goal of a typical MRC task is to require a machine to read a (set of) text passage(s) and then answer questions about the passage(s), which is a fundamental and longstanding goal of natural language understanding. MRC could be widely applied in many applications, such as, search engines, intelligent agents, dialog systems, question answering systems, and chatbots. The recent progress on the MRC task has required that the model must be capable of distinguishing those unanswerable questions to avoid giving implausible answers. MRC task with unanswerable questions may be usually decomposed into two subtasks: 1) answerability prediction and 2) reading comprehension.

So far, a common reading system which solves MRC problem generally consists of two modules: 1) building a robust language model (LM) as Encoder; 2) designing ingenious mechanisms as Decoder according to MRC task characteristics.

For the encoder, many pre-trained language models (PrLMs) such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Alexis et al., 2020), ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020), and mT5 (Linting et al., 2021) have achieved success on various natural language processing tasks and on MRC task in other languages such as English, Chinese, French, etc., which broadly play the role of a powerful encoder by capturing the contextualized sentence-level language representations. However, here, we use XLM-RoBERTa (Alexis et al., 2020) because we find that, for machine reading comprehension task, the pre-trained language models for encoders with larger models lead to better performance. We have also tried some pre-trained language models with smaller models trained on Vietnamese datasets such as phoBERT (Nguyen and Nguyen, 2020) or multilingual datasets such as mBERT (Devlin et al.,

* Main and corresponding author.
† Equal contributions.
[1] https://vlsp.org.vn/vlsp2021/eval/mrc

2019), but not good results. In addition, the tasks on which the contextualized language models are trained also have a significant impact on the performance of the MRC models. Hardware limitations were also our limitations during experimental progress with models larger than XLM-RoBERTa (Alexis et al., 2020) in this competition.

For the decoder, recent researches on a variety of problems show that jointly learning on two or more tasks produces significant performance improvements over independent models. Therefore, we use two joint models for two main tasks: answerability prediction and pure-MRC output prediction, (i.e, predicting the start and end positions). Inspired by how humans solve reading comprehension questions. We argue that when solving a reading comprehension question or any other problem, many humans with good abilities come together to solve a question (problem) that will yield a more accurate answer (result). Thus, we use ensemble models in a way that gives the top K answers for each model and a verification strategy to choose the best answer.

In summary, the notable methods we experimented in this research are as follows:

- We use two types of joint models for answerability prediction and pure-MRC prediction with/ without a dependency mechanism to learn the correlation between a start position and end position in pure-MRC output prediction;
- We show that the ensemble models yield significantly better results than without ensemble models.

The rest of the paper is organized as follows: Section 2 presents the related works, Section 3 presents the proposed approach, Section 4 presents the results, and finally, Section 5 concludes the findings and future directions.

2 Related work

The research of machine reading comprehension has attracted great interest from the NLP community in the world, as well as the Vietnamese NLP community. Some early methods, such as rule-based heuristic methods (Lynette et al., 1999), (Ellen and Michael, 2000), (Eugene et al., 2000); ranking-based BM25 (Sarrouti and El Alaoui, 2017) are inspired by calculating the similarity between two sentences with several previously published algorithms; classification-based approaches (Mueller and Thyagarajan, 2016) to find out the sentence containing the answer to the question, etc. The next trend is a variety of attention-based interactions between passage and question, such as Attention Sum (Kadlec et al., 2016), Self-matching (Wang et al., 2017), Attention over Attention (Cui et al., 2017), and Bi-attention (Seo et al., 2016). Recently, deep contextual language models have been shown effective for learning universal representations by leveraging large amounts of unlabeled data and achieving various state-of-the-art results in a series of English benchmark datasets, such as SQuAD (Pranav et al., 2016), SQuAD 2.0 (Pranav et al., 2018), and NewsQA (Trischler et al., 2017). Some prominent examples are BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Alexis et al., 2020), ALBERT (Lan et al., 2020), ELECTRA (Clark et al., 2020), and mT5 (Linting et al., 2021). For the Vietnamese language, there are several MRC datasets without unanswerable questions to evaluate reading comprehension models, such as UIT-ViQuAD (Kiet et al., 2020) for Wikipedia-based texts and UIT-ViNewsQA (Kiet et al., 2021a) for health-domain news text. Many different architectures have been experimented (Kiet et al., 2020) (Phong et al., 2021) and have shown positive results.

In our experiments, we take the XLM-RoBERTa (Alexis et al., 2020) PrLM as the backbone encoder, and jointly learn answerability prediction and pure-MRC output prediction. Then, we ensemble models to achieve the best results.

3 Proposed Approach

3.1 Dataset

UIT-ViQuAD 2.0 (Kiet et al., 2021b) combines questions in UIT-ViQuAD (Kiet et al., 2020) containing about 23K+ question-answer pairs on 170+ articles with about other 12K unanswerable questions written adversarially by crowd-workers to look similar to answerable ones. Table 1 describes details of the UIT-ViQuAD 2.0 dataset.

3.2 Our Models

Figure 1 illustrates the architecture of our two types of joint models, including an encoding layer, a decoding layer consisting of answerability prediction, and start - end position predictions (answer span prediction). In addition, we use a dependency

	Train	Public Test	Private Test	All
Number of articles	138	19	19	176
Number of passages	4,101	557	515	5173
Number of total question	28,457	3,821	3,712	35,990
Number of unanswerable questions	9,217	1,168	1,116	11,501

Table 1: *Overview of the UIT-ViQuAD 2.0 dataset.*

mechanism to enhance neural networks compared to the model without a dependency mechanism and use them for the ensemble method.

Encoder: We concatenate question and passage texts as input, which is firstly represented as embedding vectors to feed an encoding layer. The encoding layer employs a pre-trained Transformer-based language model, i.e, XLM-RoBERTa (Alexis et al., 2020) for our entire experiment. The output of the encoding layer is the contextual representations *(latest hidden state)*.

Answerability prediction: The aim of answerability prediction is to make a preliminary judgment, whether the question is answerable. Following a common strategy when fine-tuning pre-trained LMs for the sequence classification task, this layer is a linear prediction layer that is appended on top of the contextualized embedding of the classification token "[CLS]" (Devlin et al., 2019). In this experiment, we use a Softmax function instead of the usual Sigmoid for the binary classification problem because it achieves better performance. Therefore, the loss function is a Cross-entropy objective loss calculated during training.

Pure-MRC prediction: The aim of pure-MRC is to find the span in the passage as answer, i.e, find the start and end positions of that span respectively. We use the latest hidden state fed into a linear layer with Softmax operation to obtain the start and end probabilities and output the corresponding position indexes. The loss function of answer span prediction is defined as Cross-entropy for the start and end position predictions.

Dependency mechanism: The aim of this dependency mechanism is to learn the correlation between the start and end positions. We use a concatenation of the logits output of the start position prediction layer and the latest hidden state of the encoding layer through a linear layer to obtain the end position prediction.

During training, the joint loss function is the weighted sum of the answerability prediction loss (L^{ans}), start position prediction loss (L^{start}) and end position prediction loss (L^{end}):

$$L = \alpha_1 * L^{ans} + \alpha_2 * L^{start} + \alpha_3 * L^{end} \quad (1)$$

where, α_1, α_2, α_3 are weights.

3.3 Ensemble Method

The ensemble method is a combination of n models ($n = 10$ *in our best submitted results*), and a verification strategy in a way that each model gives the top K (K=20) predicted answers with corresponding predicted probabilities. For each question is a combination of all of them to aggregate for the final answer ans_{final}.

$$ans_{final} = Max(P(predicted\_answers)) \quad (2)$$

where, $P(predicted\_answers)$ is the set of probabilities of the answers that can be predicted by the models.

4 Experiments

4.1 Training Setup

The first, we build M (M=5) different datasets, each dataset comprises a train and dev set with a corresponding ratio of 0.9:0.1 that is randomly divided at the paragraph-level (article-level). For each type of model architecture, we train models and use them for ensemble models. We use the available XLM-RoBERTa (Alexis et al., 2020) PrLM - a recent state-of-the-art pre-trained language model that supports Vietnamese - as the encoder. XLM-RoBERTa is a multilingual variant of RoBERTa (Liu et al., 2019), is pre-trained on a 2.5TB multilingual dataset that contains 137GB of syllable-level Vietnamese texts.

For all experiments, we use AdamW optimizer (Ilya and Frank, 2019) with epsilon is $1e^{-8}$ and

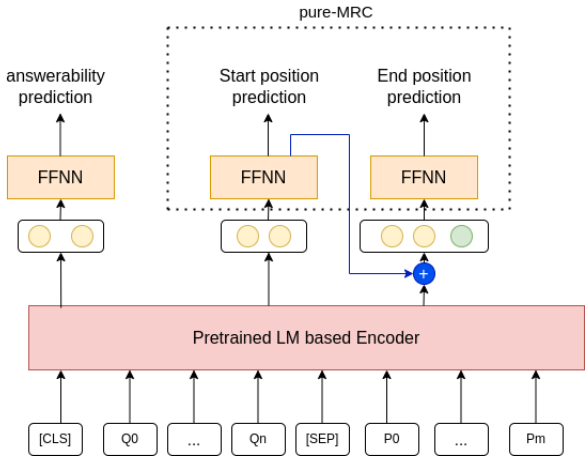

Figure 1: *Illustration of our two model types with/ without a dependency mechanism respectively with/ without blue arrow.*

learning rate to $2e^{-5}$. We set batch size to 8, max sequence length to 384, doc stride to 128, max query length to 64, and max answer length to 50. We also apply L2 weight decay with weight $1e^{-2}$. The manual weights for loss function are $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$. The maximum number of epochs is set to 10 for all experiments. All our implementations are based on the public Pytorch implementation from Transformers [2].

4.2 Evaluation Metrics

In order to more comprehensively compare the performances of MRC models, the models should be evaluated by various evaluation metrics, such as Exact Match and F1-score. These metrics are also the two metrics used in the VLSP 2021-MRC shared task competition.

Exact Match is often abbreviated as EM. If the MRC task contains *N* questions, each question corresponds to one right answer, the answers can be a word, a phrase, or a sentence, and the number of questions that the system answers correctly is *M*. Among the remaining *N* - *M* answers, some of the answers may contain some ground truth answer words, but not exactly match the ground truth answer. The Exact Match can then be calculated as follows:

$$EM = \frac{M}{N} \tag{3}$$

F1-score is a commonly used MRC task evaluation metrics. F1-score measures the overlap tokens between the predicted answers and the ground truth answers. To make the evaluation more reliable, it is also common to collect multiple correct answers to each question. Therefore, to get the average F1-score, the first has to compute the maximum F1-score of all the correct answers of a question, and then average these maximum F1-score over all of the questions. The equation of average F1-score for a task is:

$$F1 = \frac{\sum Max(F1_S)}{Num(Questions)} \tag{4}$$

where, $F1$ denotes F1-score for the MRC task, and $Max(F1_S)$ denotes the maximum F1-score of all correct answers for a single question, $\sum Max(F1_S)$ denotes the sum of for every question in the MRC task. $Num(Question)$ denotes the number of questions in the MRC task.
$F1_S$ estimated over the individual tokens in the predicted answer against those in the truth answer for a question. The equation of $F1_S$ is:

$$Precision = \frac{Num(TP)}{Num(TP) + Num(FP)} \tag{5}$$

$$Recall = \frac{Num(TP)}{Num(TP) + Num(FN)} \tag{6}$$

$$F1_S = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

where, for a single question, the token-level true positive (TP) denotes the same tokens between the predicted answer and the truth answer. The token-level false positive (FP) denotes the tokens which are not in the truth answer but the predicted answer, while the false negative (FN) denotes the tokens which are not in the predicted answer but the truth answer.

[2] https://huggingface.co/transformers/index.html

4.3 Results

Table 2 shows the performance of our method and the baseline on the public and private test sets of the UIT-ViQuAD 2.0 dataset. Our model achieves 80.578% (in $F1 - score$), 70.662% (in EM) on public test dataset and achieves 76.456% (in $F1 - score$), 64.655% (in EM) on private test dataset. Compared to the baseline model, our model achieves much better results.

Method	Public Test		Private Test	
	F1	EM	F1	EM
Our model	**80.578**	**70.662**	**76.456**	**64.655**
Baseline	63.031	53.546	60.338	49.353

Table 2: *The results on the public and private test sets of the UIT-ViQuAD 2.0 dataset, evaluated on EM and F1 scores.*

Table 3 compares the leading model with/ without ensemble models on the public test of the UIT-ViQuAD 2.0 dataset. The model with ensemble models (our model) outperformed the model without ensemble models only achieving 76.657% (in $F1 - score$) and 65.768% (in EM).

Method	Public Test	
	F1	EM
With Ensemble	**80.578**	**70.662**
Without Ensemble	76.657	65.768

Table 3: *The results on the public test sets of the UIT-ViQuAD 2.0 dataset, comparisons between with/ without ensemble models.*

5 Conclusion

In this paper, we describe and propose our approach to solve the Vietnamese Machine Reading Comprehension competition in the evaluation campaign of the VLSP 2021-MRC Shared Task. For the future, we would like to experiment some pre-trained language models with different and larger architectures such as GPT (Tom et al., 2020), mT5 (Linting et al., 2021), etc., because MRC systems greatly benefit from the development of pre-trained language models and simultaneously research some other verification strategies to improve model performance.

Acknowledgments

We would like to thank the VLSP 2021 organizers for their really hard work in maintaining the conference every single year and providing the dataset of Vietnamese Machine Reading Comprehension for our experiments.

References

Conneau Alexis, Khandelwal Kartikay, Goyal Naman, Chaudhary Vishrav, Wenzek Guillaume, Guzmán Francisco, Grave Edouard, Ott Myle, Zettlemoyer Luke, and Stoyanov Veselin. 2020. Unsupervised cross-lingual representation learning at scale. *arXiv preprint*, "arXiv:1911.02116".

K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. Electra: Pre-training text encoders as discrimi- nators rather than generators. *In International Conference on Learning Representations*, "arXiv:2003.10555".

Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu. 2017. Attention-over-attention neural networks for reading comprehension. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, "arXiv:1810.04805".

Riloff Ellen and Thelen Michael. 2000. A rule-based question answering system for reading comprehension tests. *In Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems*.

Charniak Eugene, Altun Yasemin, de Salvo Braz Rodrigo, Garrett Benjamin, Kosmala Margaret, Moscovich Tomer, Pang Lixin, Pyo Changhee, Sun Ye, Wy Wei, Zhongfa Yang, Zeiler Shawn, and Zorn Lisa. 2000. Reading comprehension programs in a statistical-language-processing class. *In Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems*.

Loshchilov Ilya and Hutter Frank. 2019. Decoupled weight decay regularization. *In Proceedings of ICLR*, "arXiv:1711.05101".

R. Kadlec, M. Schmid, O. Bajgar, and J. Kleindienst. 2016. Text understanding with the attention sum reader network. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Van Nguyen Kiet, Vu Nguyen Duc, Gia-Tuan Nguyen Anh, and Luu Thuy Nguyen Ngan. 2020. A vietnamese dataset for evaluating machine reading comprehension. *The 28th International Conference on Computational Linguistics*, arXiv:2009.14725.

Van Nguyen Kiet, Nguyen Duc-Vu, Gia-Tuan Nguyen Anh, and Luu-Thuy Nguyen Ngan. 2021a. New vietnamese corpus for machine reading comprehension of health news articles. In arXiv:2006.11138.

Van Nguyen Kiet, Quoc Tran Son, Thanh Nguyen Luan, Van Huynh Tin, Son T. Luu, and Luu-Thuy Nguyen Ngan. 2021b. Vlsp 2021 shared task: Vietnamese machine reading comprehension. *In Proceedings of the 8th International Workshop on Vietnamese Language and Speech Processing (VLSP 2021)*.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. A lite bert for self- supervised learning of language representations. *In International Conference on Learning Representations.*, "arXiv:1909.11942".

Xue Linting, Constant Noah, Roberts Adam, Kale Mihir, AlRfou Rami, Siddhant Aditya, Barua Aditya, and Raffel Colin. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint*, "arXiv:2010.11934".

M. Liu, Y.and Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*, "arXiv:1907.11692".

Hirschman Lynette, Light Marc, Breck Eric, and D. Burger John. 1999. Deep read: A reading comprehension system. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 189–198.

J. Mueller and A. Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. *In: Proceedings of the AAAI Conference on Artificial Intelligence*, 30:2786–2792.

D.Q. Nguyen and A.T. Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *In Findings of the Association for Computational Linguistics: EMNLP*, arXiv:2003.00744.

Nguyen-Thuan Do Phong, Duy Nguyen Nhat, Van Huynh Tin, Van Nguyen Kiet, Gia-Tuan Nguyen Anh, and Luu-Thuy Nguyen Ngan. 2021. Sentence extraction-based machine reading comprehension for vietnamese. *In International Conference on Knowledge Science, Engineering and Management*, arXiv:2105.09043.

Rajpurkar Pranav, Zhang Jian, Lopyrev Konstantin, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint*, arXiv:1606.05250.

Rajpurkar Pranav, Jia Robin, and Liang Percy. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint*, arXiv:1806.03822.

M. Sarrouti and S.O. El Alaoui. 2017. A passage retrieval method based on probabilistic information retrieval model and umls concepts in biomedical question answering. *Journal of biomedical informatics*, 68:96–103.

M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *In International Conference on Learning Representations*, arXiv:1611.01603.

B. Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, M. Ziegler Daniel, Wu Jeffrey, Winter Clemens, Hesse Christopher, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, McCandlish Sam, Radford Alec, Sutskever Ilya, and Amodei Dario. 2020. Language models are few-shot learners. *arXiv preprint*, arXiv:2005.14165.

A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, Bachman P., , and K. Suleman. 2017. Newsqa: A machine comprehension dataset. *In Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 189–198.