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: Context-Aware Answer Extraction in Vietnamese Question Answering

Le Thi Thu Hang*, Ho Duc Viet, Nguyen Duc Vu

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

Received 27 December 2021
Revised 30 March 2022; Accepted 4 May 2022

Abstract: Machine Reading Comprehension is one of the natural language processing fields; machines automatically have to answer questions based on specific passages for this task. In recent years, machine reading comprehension (MRC) has received much attention; many articles have been written about this task. However, most articles only develop models in two main languages, English and Chinese. In this article, we propose to apply a new model to the task of reading comprehension in Vietnamese. Specifically, we use BLANC (BLock AttentioN for Context prediction) on pre-trained baseline models to solve the Machine reading comprehension (MRC) task on Vietnamese. We have achieved good results when using BLANC on the baseline model. Specifically, with the MRC task at the VLSP Share-task 2021, we scored 77.222% of F1-score on the private test and ranked 2nd in the total. This shows that BLANC method works very well in MRC tasks and further enhances the Vietnamese MRC development.

Keywords: VLSP 2021, Answer Extraction, BLANC.

1. Introduction

Machine Reading Comprehension (MRC) is the fundamental and long-term goal of natural language understanding (NLU). It aims to teach the machine to understand the given passage and answer questions based on it. MRC has many vital applications, such as automatic dialog and question answering systems. Question answering tasks require a high level of reading comprehension, resulting to high requirements of language understanding. So that is the reason the language models use question-answering (QA) tasks to evaluate various language comprehension tasks.

The original MRC systems were designed with the assumption that all questions could be answered according to the given passage, which is not always the case in the real world. The recent progress of the MRC task has required that the model distinguishes between unanswerable questions and answerable questions to avoid giving unreasonable answers. To solve this challenge, the model must carefully

* Corresponding author.
E-mail address: 18520274@gm.uit.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.316

handle two aspects: i) distinguish unanswerable questions effectively; ii) give correct answers to answerable questions.

Table 1. An example is that the correct answer appears multiple times. The first occurrence of "cà phê" is the correct answer; the remaining words are incorrect

Passage: Việt Nam là một nước xuất khẩu cà phê, do đó nhiều loại cà phê được sử dụng ngày càng thịnh hành trong ẩm thực của người Việt tại khắp các vùng miền, đặc biệt tại cácđô thị. Cà phê thường được pha, chiết bằng phin pha cà phê. Theo thuộc tính nhiệt, có thể kể ra hai cách uống phổ biến là cà phê nóng và cà phê đá, xét theo nguyên liệu phụ gia, cà phê thuần nhất gọi là cà phê đen, và cà phê sữa. Nước chiết cà phê cũng thường dùng đểchế thêm vào một số loại nước sinh tố hay sữa chua cho hương vị đặc biệt. Ngày nay, càphê hòa tan cũng là loại cà phê thông dụng. *(Vietnam is a coffee exporter, so many typesof coffee are used more and more popularly in Vietnamese cuisine in all regions, especiallyin urban areas. Coffee is usually brewed and extracted by a coffee filter. According to thermal properties, two popular ways of drinking can be mentioned as hot coffee and iced coffee. According to the additive ingredients, pure coffee is called black coffee, and milk coffee. Coffee extract is also often used to make some smoothies or yogurts for a special flavor. Nowadays, instant coffee is also popular coffee.)*
Question: Việt Nam được biết đến là một trong những nước xuất khẩu gì? *(Whatis Vietnam known as one of the exporting countries?)*
Answer: cà phê *(coffee)*

In the International Workshop on Vietnamese Language and Speech Processing (VLSP) 2021: the Vietnamese Machine Reading Comprehension task [1] required participants to build the Vietnamese MRC model from the given dataset UIT-ViQuAD2.0. It combined answerable questions from the UIT-ViQuAD1.0 dataset [2] and unanswerable questions about the same passage. To work effectively on this dataset, the MRC model must answer the answerable questions and identify questions that cannot be answered. For the issue one mentioned above, get ideas from Retrospective Reader [3]; in addition to the in-depth question answering reader, we add a task classifier for unanswerable classification. Moreover, to predict accurate answers to answerable questions, we researched and decided to apply BLANC to our advanced reader.

According to the published BLANC article [4], using BLANC to help increase the readability of the model resulted in an accuracy increase of 2-3% compared to the baseline models on English datasets. The baseline models will extract the answer with the highest probability from the given passage. However, they sometimes predict the correct answer but the contexts unrelated to the given question. This distinction becomes especially important as the number of occurrences of the answer in a passage increases. Table 1 shows an example of the correct answer that appears multiple times in the passage 1. An overall illustration of BLANC [4] suggested BLANC to overcome this challenge, which is built on two primary ideas: utilizing a block attention method to learn the soft-labels and using the soft-labels method to predict the context of the question. Specifically, we will present more details in section 3.2.

In this paper, we have three main contributions as follows:

• Firstly, we implemented two models based on transformers such as XLM-RoBERTa, Rembert to solve the VLSP2021 MRC task.

• Secondly, we propose to apply BLANC to the model to test the performance of BLANC on the Vietnamese MRC model.

• Finally, we ensemble Rembert and XLM Roberta models after using BLANC, and we achieved positive results in the Vietnamese MRC task.

The paper's organization is as follows: In section 2, we will discuss some related works on

this topic, and in section 3, we will explain more about our methodology. Section 4 is our experiment and results, and section 5 is the discussion. Section 6 is the conclusion and the future work.

2. Related Work

Machine Reading Comprehension has attracted significant attention from researchers, with the release of many task-specific datasets [5, 6, 7]. Along with that, many studies on MRC have been developed, researchers initially focused on attention-based interactions between passages and questions, including Attention Sum [8], Gated attention [9], Self-matching [10], Attention over Attention [11] and Biattention [12]. Recently, many powerful PrML models such as ELMo [13], GPT [14], BERT [15], XLNet [16], RoBERTa [17], ALBERT [18] have been successfully used in machine reading comprehension.

MRC is also starting to gain attention in Vietnam by releasing the specific datasets in Vietnamese such as ViQuAD1.0 [2], ViMMRC [19], UIT-ViNewsQA [20]. In addition, many research articles related to machine learning understanding Vietnamese have been published, some typical works like [21, 22, 23, 24] showing the development of Vietnamese MRC. Later, with higher model readability requirements, later versions of the datasets added unanswered questions. It is important to solve the MRC task with the unanswerable question, but few studies focus on this topic. Inspired by the retrospective reader [3] that integrates two reading and verification strategies stages, our model structure consists of a reading module to predict the answer and another reading module to classify unanswerable answers.

The intensive reading module receives interest with many articles published and many ideas presented [25, 26]. These ideas and models have shown promising performance in predicting the answer to a given question and passage. However, most previous research did not focus on giving answers in the context of the question. BLANC [4] is proposed that solved this problem, and we applied it to our reading module.

3. Methodologies

Our model uses three main phases: 1) Data preprocessing; 2) Reading module; 3) Output processing. Specifically, we will describe and present it below.

3.1. Dataset Preprocessing

Trankit: Trankit [27] is a light-weight Transformer-based Toolkit for multilingual Natural Language Processing (NLP) that provides trainable pipelines over 100 languages for some fundamental NLP tasks. Trankit can process inputs that are untokenized (raw) or pre-tokenized strings at both sentence and document levels. Currently, Trankit supports the following tasks: sentence segmentation, tokenization, part-of-speech tagging, and so on. We use tokenization to separate punctuation and align the segmented answers in this task.

Extract the sentence containing the answer: To effectively make the model distinguish between answerable and unanswerable questions, we generate a new dataset from the original train dataset. This dataset is dedicated to determining whether a question can be answered or cannot be answered. We make the new dataset this way: for the question to be answered, we rely on the start and end answers index to extract the one or more sentences containing the answer in the passage; for the unanswerable, we use plausible answers instead of answers.

3.2. Reading Module

This section presents the architecture of our BERT-base models that we used, the BLANC with Rembert [28] and XLM-RoBERTa [29].We will present the reasons for choosing thesetwo pretrained models in section 5. RemBERT is considered a larger version of the BERT multilingual with decoupled input and output embedding, which speeds up the training process. RemBERT has been pre-trained on large unlabeled text using Wikipedia and

Common Crawl data, covering 110 languages. XLM-RoBERTa is the multilingual version of Roberta, a cross-lingual transformer pre-trained on Wikipedia text in 100 languages, and it has been highly influential in many tasks. block attention method to predict the answer in the context of a given question when the passage has the answer texts two or more. BLANC uses the soft-labeling method and

Figure 2 shows the BLANC's architecture. Soft-labeling for latent context C: The authors calculate the probability of a word being in the context of a given question, $p_{soft}(w_i \in C)$, used for the context prediction task. To achieve this, the authors hypothesized that the farther the words are from the answer, the lower the probability of being in context C decreases by a certain ratio q.

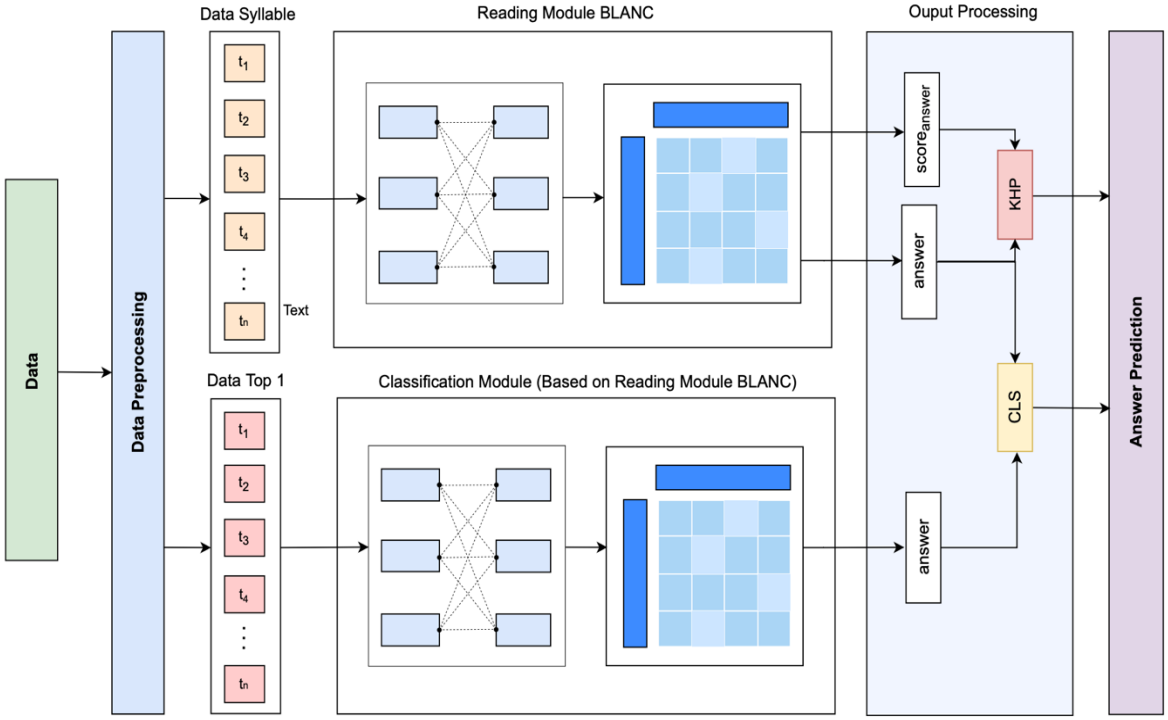

Figure 1. The visualization of our model. 1) Process the original dataset using trankit, and make a new dataset forthe classification module. 2) Use 10-fold cross-validation on two datasets to train the model. 3) At the output processing stage, ensemble all models after training and deal with the duplicated answers.

In addition, for computational efficiency, the authors employ the hyper-parameter window size to be bound on both sides of an answer-span, which means that $p_{soft}(w_i \in C)$ equals 0 when the word is outside the bounds of window size. Block attention: Specifically, the task of this stage is to calculate $p(w_i \in C)$ to predict the soft label and the index of answer-span. We compute the probability $p(w_i \in C)$ in the following steps:

i) predicting the context span containing answer, $p(i = s_c)$ and $p(i = e_c)$, and

ii) calculating $p(w_i \in C)$ using the cumulative distributions of $p(i = s_c)$ and $p(i = e_c)$.

By reducing the cross-entropy of the two probabilities, $p(w_i \in C)$ and $p_{soft}(w_i \in C)$, we explicitly require the block attention model to learn context words of a given question. The

following equation defines the latent context's loss function:

$$L_{context} = -\sum_{1\le i\le l} p_{soft}(w_i \in C)\log p(w_i \in C) - \sum_{1\le i\le l} p_{soft}(w_i \notin C)\log p(w_i \notin C) \quad (1)$$

where l is the length of a passage. The final context loss function is obtained by averaging Lcontext across all train samples.

Answer-span Prediction: For the answer-span prediction layer as BERT, instead of simply multiplying the weights by the encoder output, the authors also multiplied the context probability p(wi C) calculated in the previous section to pay more attention to the answer-span in the context, C.

The following equation describes the loss function for answer-span prediction:

$$L_{answer} = -\frac{1}{2}\{\sum_{1\le i\le l} \mathbb{I}(i = s_a)\log p(i = s_a) + \sum_{1\le i\le l} \mathbb{I}(i = e_a)\log p(i = e_a)\} \quad (2)$$

l is a conditional function that returns 1 when the condition is true and 0 otherwise. The authors average the Lanswer over all train samples as the final answer-span loss function and take the weighted sum of the two loss functions as the loss function of our final prediction:

$$L_{total} = (1-\lambda)L_{answer} + \lambda L_{context} \quad (3)$$

where λ is a hyper-parameter modifies the loss function ratio of two loss functions.

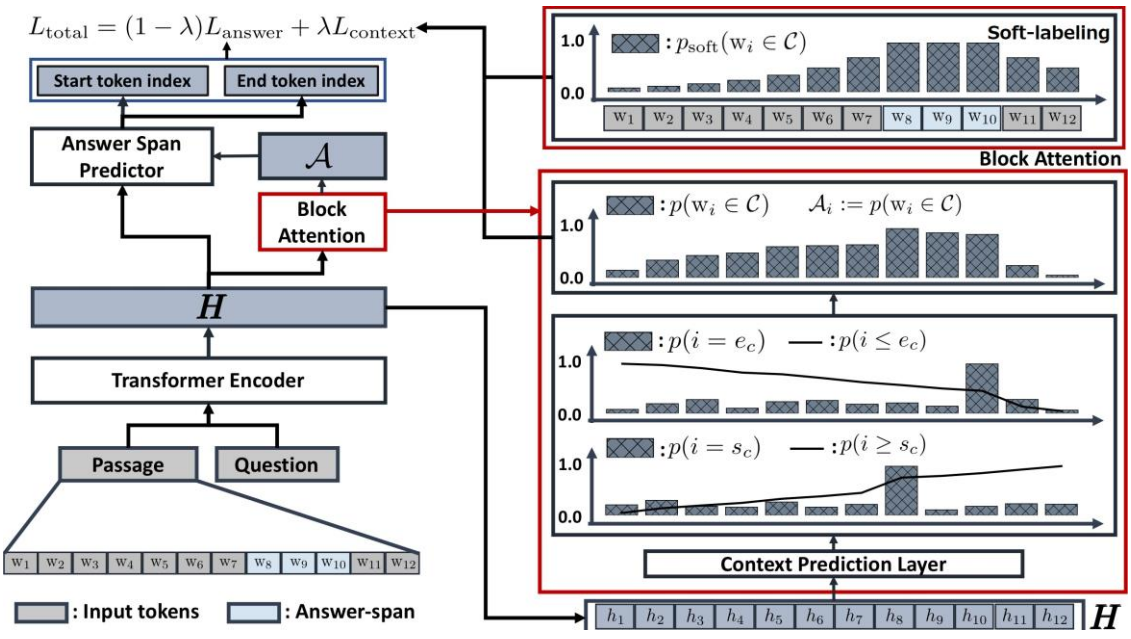

Figure 2. An overall illustration of BLANC [4].

3.3. Output Processing

Before giving the final result, we have processed the output to achieve higher results. Specifically, we have ensemble two models, RemBERT and XLM-R. Then, we applied the keeping the highest probability or the classification module to guess the unanswerable questions from the dataset containing duplicated answer questions that we filtered out.

Ensemble: We calculate the average probability of the answer prediction from models; then, we choose the highest one. We ensemble XLM-RoBERTa and SemBERT models with a 1:1 ratio.

Keep the highest probability among duplicate answer (HP): We use the scoreanswer, the probability of the answer predicted by the advanced reader, to choose the answer with the highest probability among duplicated predictions answers but have different questions as the final answer; the ones with lower probability become unanswerable.

Classification Module: We trained this model as a reading model on the top 1 dataset we created previously and used it to specialize in classifying unanswerable questions; this model also predicts the probability of an answer. Instead of using HP, we will use this model to classify duplicated answer predictions with different questions.

In submission 2 we keep the highest probability among duplicate answers as output processing, and in submission 3 we used the classification module to process output.

Table 2. Some information about the UIT-ViQuAD2.0 dataset

	Train	**Public Test**	**Private Test**	**All**
Number of articles	138	19	19	176
Number of passages	4101	557	515	5173
Number of total questions	28,457	3,821	3,712	35,990
Number of unanswerable questions	9,217	1,168	1,116	11,501

4. Experiments and Results

4.1. Dataset

UIT-ViQuAD [2] stands for Vietnamese Question Answering Dataset, which was created specifically for the task of Vietnamese machine reading comprehension based on passages extracted from Vietnamese Wikipedia articles. Initially, UIT-ViQuAD1.0 consisted of only answerable questions. To increase machine learning, the UIT-ViQuAD2.0 [1] combined 23K questions in UIT-ViQuAD1.0 with more than 12K unanswered questions.

We evaluated our method on the UIT-ViQuAD2.0 datasets provided by the VLSP Shared Task 2021, containing over 35K questions. Table 2 is some information about UIT-ViQuAD2.0.

4.2. Evaluation Metrics

The VLSP Share-task 2021 is evaluated and ranked using the standard evaluation metrics in the MRC, which are F1-score and EM score. The final ranking is evaluated according to the F1-Score; EM is a secondary metric.

Exact Match (EM) is the number of precise answers, giving a score of 1 when the prediction and the true answer are the same and 0 otherwise. When evaluating against a negative question, if the system predicts any textual span as an answer, it automatically obtains a zero score for that question.

F1-score estimated over the individual tokens in the predicted answer against those in the gold standard answers is based on the number of matched tokens between the expected and gold standard answers.

$$Precision = \frac{N_{matched}}{N_{predicted}}$$

$$Recall = \frac{N_{matched}}{N_{gold\_standard}}$$

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall}$$

where $N_{matched}$ is the number of matched tokens. $N_{predicted}$ and $N_{gold\_standard}$ is the numberof the predicted answer tokens and the gold standard answer tokens respectively.

4.3. Model Settings

Using a 10-fold cross-validation approach, we divided the dataset into a training set and a validation set, for both the original and top 1 datasets. For each fold, we use AdamW [30] for optimization with a learning rate of 2 10−5 and batch size of 4 for XLM-RoBERTa and 2 for RemBERT. Warm-up learning was applied, with a maximum learning rate of 0.1 selected. The embedding parameters of the model are as follows: max_seq_length=384, max_query_ length=128, max_answer_length = 500 and doc_stride = 128. We trained our model on 3 epochs and used a cross-entropy as the loss function. In addition, we use 3 parameters of BLANC, q = 0.7, window-size = 2 and $\lambda = 0.4$.

4.4. Results

In the VLSP Share-task 2021 with machine reading comprehension, we applied BLANC for XLM-Roberta and RemBERT on a 10-fold cross-validation dataset. Then, we ensemble the trained models to get the best answer. After that, we choose the highest probability prediction answer among the duplicated answers that we filtered from the results of the previous reading module. The result was 77.222% f1-score (top 2) and 67.430% EM (top1) on the private test, resulting in the second submission and the highest one. Table 3 shows the leaderboard on the private test on this Share-task.

Table 3. The leaderboard on private test

Teams	**F1**	**EM**
vc-tus	77.241	66.137
ebisu_uit	**77.222**	**67.430**
F-NLP	76.456	64.655

In submission 3, instead of selecting the highest result among a duplicated answer as submission 2, we use the classification module to classify those answers. Still, the result is not as good as the second submission. Table 4 shows our result submission on private test

Table 4. Our submission on private test

	F1	**EM**
Submission 2	**77.222**	**67.430**
Submission 3	76.877	66.676

5. Discussion

Table 5. The result using BLANC on some pre-trained models

Fold	**BLANC**			
	Pho-BERT	BART-pho	Rem-BERT	XLMR
1	67.898	66.934	73.217	74.360
2	71.453	67.944	74.486	75.478
3	70.296	65.682	73.194	74.428
4	71.543	65.077	74.572	74.533
5	69.020	64.587	71.854	71.967
Avg	70.042	66.045	**73.465**	**74.153**

Based on results using BLANC on some pre-trained models in table 5, the two multilingual RemBERT and XLM-RoBERTa models can be seen as better results than the two Vietnamese monolingual BartPho and PhoBERT models. This can be explained by the long passages in the dataset, whereas the PhoBERT monolingual model only supports tokens with a maximum length of 256, which leads to the model having to truncate many times, leading to loss of context. It does not happen for RemBERT and XML-RoBERTa, which supports max length tokens up to 512, twice as much as PhoBERT.

We have implemented to prove that the max token length of the model affects the model's results, the f1-score of XLMR with 256 tokens is lower than that of 512 tokens, detailed in table 6.

The BARTPho model has achieved quite good results in the task of natural language understanding, precisely the Vietnamese text summarization problem. However, BARTPho does not achieve good results compared to other multilingual models in this problem. One of the reasons is that BARTpho's architecture is based on BART's architecture, a model created to specialize in processing sequence to sequence problems that should not work well for this problem.

Then, we present our experiment using BLANC and not using BLANC on two models,

XLM-RoBERTa and RemBERT, using the UIT-ViQuAD2.0 dataset with a 5-fold cross-validation method. Table 7 shows our results of using BLANC and not using BLANC. The model using BLANC increased about 0.2 - 0.4% compared to the baseline model, giving better results but not much.

Table 6. The effect of max token length on model results

Fold	XMLR - 256	XMLR - 512
1	71.768	74.360
2	72.844	75.478
3	71.702	74.428
3	72.555	74.533
5	69.677	71.967
Avg	71.709	**74.153**

Table 7. The result of using BLANC and not using BLANC

Fold	without BLANC		BLANC	
	Rem	XLMR	Rem	XLMR
1	72.936	74.051	73.217	74.360
2	74.561	74.936	74.486	75.478
3	72.643	73.478	73.194	74.428
4	74.071	74.942	74.572	74.533
5	72.036	71.568	71.854	71.967
Avg	73.249	73.795	**73.465**	**74.153**

According to the authors, the BLANC model will work well on datasets in which the answer text appears more than once in the passage. On the UIT-ViQuAD2.0 dataset, the question with more than one answer text in the passage accounts for about 8.5% of the total dataset (figure 3). Therefore, BLANC does not work effectively on this dataset, but BLANC still obtains better results than conventional baseline models. We expect the model to get better results in later developed datasets.

When our model made predictions on the validation set, we found that the model predicts many of the same answers but different questions; and only 1-2 of those answers were correct. That could be because the MRC module does not work well in predicting answers to complex questions. Therefore, we propose two solutions to solve the above problem: keep the highest probability among duplicate answers method and the classification module (in BLANC MRC format) that we introduced above.

Table 8 is the result of keeping the answer with the highest probability among the duplicate answers. And the obtained results are 2% - 3% better than the original. Table 9 shows the influence of the classification module on the results. Specifically, when using the classification module, the results increased by about 1.5%-2%. The two proposed solutions have obtained quite positive results for predicting answers to complex questions in the dataset.

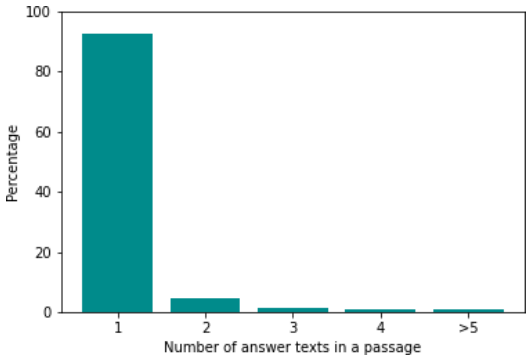

Figure 3. Proportions of questions with various numbersof the answer text in a passage

On the UIT-ViQuAD2.0 dataset, the question with more than one answer text in the passage accounts for about 8.5% of the total dataset (figure 3). Therefore, BLANC does not work effectively on this dataset, but BLANC still obtains better results than conventional baseline models. We expect the model to get better results in later developed datasets.

When our model made predictions on the validation set, we found that the model predicts many of the same answers but different questions; and only 1-2 of those answers were correct. That could be because the MRC module does not work well in predicting answers to complex questions. Therefore, we propose two solutions to solve the above problem: keep the highest probability among duplicate answers method and the classification module (in BLANC MRC format) that we introduced above.

Table 8 is the result of keeping the answer with the highest probability among the duplicate

answers. And the obtained results are 2% - 3% better than the original. Table 9 shows the influence of the classification module on. the results. Specifically, when using the classification module, the results increased by about 1.5%-2%. The two proposed solutions have obtained quite positive results for predicting answers to complex questions in the dataset.

Table 8. The result of keeping the highest probability answer among the duplicate answers

Fold	BLANC		BLANC + HP	
	Rem	XLMR	Rem	XLMR
1	73.217	74.360	76.368	76.530
2	74.486	75.478	77.849	78.133
3	73.194	74.428	75.749	75.873
4	74.572	74.533	77.722	77.423
5	71.854	71.967	74.741	74.660
Avg	73.465	74.153	**76.486**	**76.524**

Table 9. The influence of the classification module

Fold	BLANC		BLANC + CLS	
	Rem	XLMR	Rem	XLMR
1	73.217	74.360	74.321	75.560
2	74.486	75.478	75.769	77.398
3	73.194	74.428	74.735	75.259
4	74.572	74.533	76.526	76.078
5	71.854	71.967	73.240	73.830
Avg	73.465	74.153	**74.918**	**75.625**

6. Conclusion

6.1. Summary

There are some of our contributions in this paper:

• We applied BLANC to the model to test the performance of BLANC on the Vietnamese MRC model.

• We ensembled two models, XLM-R and RemBERT, to get good performance.

• We proposed two solutions to deal with difficult questions in the dataset.

We achieved good results on the MRC task at the VLSP share-task 2021 with a 77.222% f1-score (top 2) and 67.430% EM (top 1).

6.2. Future Work

After rechecking the passages and the answers, we found that in creating the new dataset through extracting the sentences containing the answers, we only get 1 to 2 sentences containing the answers. That leads to the loss of the passage's context, so the classification model's training and prediction are affected. We will increase the number of sentences retrieved to get the full context of the answer in the future.

Due to the time limit, we have not been able to adjust and choose the suitable parameters for BLANC, so we will try to adapt and test many other parameters to choose the most relevant parameters for BLANC. And we will try to apply some more methods like Named Entity Recognition (NER) and Dependency Parsing to improve the results.

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

[1] Nguyen Van Kiet, Tran Quoc Son, Nguyen Thanh Luan, Huynh Van Tin, Luu T. Son and Nguyen Luu-Thuy Ngan. 2022. VLSP 2021 - ViMRC Challenge: Vietnamese Machine Reading Comprehension. VNU Journal of Science: Computer Science and Communication Engineering, 38(2).

[2] 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.

[3] Z. Zhang, J. Yang, H. Zhao, Retrospective reader for machine reading comprehension, arXiv preprint arXiv:2001.09694.

[4] Y. Seonwoo, J.-H. Kim, J.-W. Ha, A. Oh, Context-Aware Answer Extraction in Question Answering, in: Proceedings of the 2020 Conference on

Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2418–2428.

[5] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1601–1611.

[6] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2383–2392.

[7] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, RACE: Large-scale Reading Comprehension Dataset From Examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 785–794.

[8] R. Kadlec, M. Schmid, O. Bajgar, J. Kleindienst, Text Understanding With The Attention Sum Reader Network, arXiv:1603.01547.

[9] B. Dhingra, H. Liu, Z. Yang, W. W. Cohen, R. Salakhutdinov, Gated-Attention Readers for Text Comprehension, arXiv:1606.01549.

[10] W. Wang, N. Yang, F. Wei, B. Chang, M. Zhou, Gated Self-Matching Networks for Reading Comprehension and Question Answering, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2017, pp. 189–198.

[11] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, G. Hu, Attention-Over-Attention Neural Networks For Reading Comprehension, arXiv:1607.04423.

[12] M. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, Bidirectional Attention Flow For Machine Comprehension, arXiv:1611.01603.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.

[14] A. Radford, K. Narasimhan, T. Salimans, Sutskever, Improving language understanding by generative pre-training. https://gluebenchmark.com/leaderboard. Accessed in October 2021.

[15] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, 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, Vol. 1, 2019, pp. 4171–4186.

[16] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances In Neural Information Processing Systems, Vol. 32, 2021, pp.134-141.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A Robustly Optimized Bert Pretraining Approach, arXiv:1907.11692.

[18] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A Lite Bert for Self-Supervised Learning of Language Representations, arXiv:1909.11942.

[19] 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.

[20] 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, arXiv:2006.11138.

[21] K. V. Nguyen, N. D. Nguyen, P. N. T. Do, A. G. T. Nguyen, N. L. T. Nguyen, ViReader: A Wikipedia-based Vietnamese reading comprehension system using transfer learning, Journal of Intelligent & Fuzzy Systems , 2021, pp1–19.

[22] 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, in: International Conference on Computational Collective Intelligence, Springer, 2021, pp. 546–558.

[23] 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, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2021, pp. 511–523.

[24] S. T. Luu, K. V. Nguyen, A. G. T. Nguyen, N. L. T. Nguyen, An Experimental Study of Deep Neural Network Models for Vietnamese Multiple Choice Reading Comprehension, in: 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE), IEEE, 2021, pp. 282–287.

[25] K. Sun, D. Yu, D. Yu, C. Cardie, Improving Machine Reading Comprehension with General Reading Strategies, arXiv:1810.13441.

[26] Y. Dai, Y. Fu, L. Yang, A Multiple-Choice

Machine Reading Comprehension Model with Multi-Granularity Semantic Reasoning, Applied Sciences, Vol. 11, No. 17, 2021, pp. 7945.

[27] M. V. Nguyen, V. D. Lai, A. P. B. Veyseh, T. H. Nguyen, Trankit: A Light-Weight Transformer-Based Toolkit for Multilingual Natural Language Processing, arXiv:2101.03289.

[28] H. W. Chung, T. Févry, H. Tsai, M. Johnson, S. Ruder, Rethinking Embedding Coupling in Pre-Trained Language Models, arXiv:2010.12821.

[29] 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, arXiv:1911.02116.

[30] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, arXiv:1711.05101.