VNU Journal of Science:
Computer Science and Communication Engineering

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

Original Article

VLSP 2021-ViMRC Challenge: Vietnamese Machine Reading Comprehension

Nguyen Van Kiet[1,2], Tran Quoc Son[3], Nguyen Thanh Luan[1,2], Huynh Van Tin[1,2], Luu Thanh Son[1,2], Nguyen Luu Thuy Ngan[1,2,*]

[1]*University of Information Technology, Vietnam National University, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City Vietnam*
[2]*Vietnam National University, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City Vietnam*
[c]*Denison University, Granville, OH, USA*

Received 27 December 2021
Revised 18 April 2022; Accepted 5 May 2022

Abstract: One of the emerging research trends in natural language understanding is machine reading comprehension (MRC) which is the task to find answers to human questions based on textual data. Existing Vietnamese datasets for MRC research concentrate solely on answerable questions. However, in reality, questions can be unanswerable for which the correct answer is not stated in the given textual data. To address the weakness, we provide the research community with a benchmark dataset named UIT-ViQuAD 2.0 for evaluating the MRC task and question answering systems for the Vietnamese language. We use UIT-ViQuAD 2.0 as a benchmark dataset for the challenge on Vietnamese MRC at the Eighth Workshop on Vietnamese Language and Speech Processing (VLSP 2021). This task attracted 77 participant teams from 34 universities and other organizations. In this article, we present details of the organization of the challenge, an overview of the methods employed by shared-task participants, and the results. The highest performances are 77.24% in F1-score and 67.43% in Exact Match on the private test set. The Vietnamese MRC systems proposed by the top 3 teams use XLM-RoBERTa, a powerful pre-trained language model based on the transformer architecture. The UIT-ViQuAD 2.0 dataset motivates researchers to further explore the Vietnamese machine reading comprehension task and related tasks such as question answering, question generation, and natural language inference.

Keywords: Machine Reading Comprehension, Question Answering, VLSP, Transfer Learning.

1. Introduction

Machine Reading Comprehension (MRC) is an emerging and challenging task of natural language understanding that computers can read and understand texts and then find correct answers to any questions. Recently, many MRC shared tasks [1-3] and benchmark corpora [4-11]

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

https://doi.org/10.25073/2588-1086/vnucsce.340

have attracted a range of researchers from academia and industry. Therefore, significant progress has been exploited in building computational models for semantics based on deep neural networks and transformers over the last ten years [12-15]. The datasets and models are studied in resource-rich languages such as English and Chinese. Until now, no MRC shared task or challenge has been organized for Vietnamese, which motivates us to organize a challenge for Vietnamese machine reading comprehension.

We introduce the VLSP 2021-ViMRC Challenge: Vietnamese Machine Reading Comprehension. We hope to use this challenge to examine the capabilities of state-of-the-art deep learning and transformer models to represent and simulate machine reading comprehension for Vietnamese texts. Inspired by machine reading comprehension benchmarking [6], we design this challenge of Vietnamese reading comprehension, in which computers are given a document D as well as a human question Qi to comprehend. In this work, we construct UIT-ViQuAD 2.0, a new dataset that combines answerable questions from the previous version of UIT-ViQuAD (UIT-ViQuAD 1.0 [16]) with over 12K unanswerable questions for the same passages. Figure 1 illustrates two such examples.

The participating teams made 590 total submissions within the official VLSP-2021 evaluation period. We introduce the challenge and present a summary for the evaluation in this paper.

In this paper, we have three main contributions described as follows.

• Firstly, we constructed UIT-ViQuAD 2.0, a Vietnamese dataset for the span-extraction reading comprehension task which contains nearly 36,000 human-annotated questions including unanswerable and answerable. Unanswerable questions are added to increase the linguistic diversity in machine reading comprehension and question answering.

• Secondly, we organize the VLSP 2021-ViMRC Challenge for evaluating MRC and question answering models in Vietnamese at the VLSP 2021. Our baseline system obtains 63.03% and 60.34% in F1-score on the public and private test sets, respectively, and there is no model of participating teams that pass 78% (in F1-score) on the private test set, which indicates our dataset is challenging and requires the development of MRC models for the Vietnamese.

• UIT-ViQuAD 2.0 could also be a good resource for multilingual and crosslingual research purposes when studied along with other MRC and QA datasets.

The following is how the rest of the article is organized. In Section 2, we provide a brief overview of the background and relevant studies. We introduce the VLSP 2021-ViMRC Challenge in Section 3. Our new dataset (UITViQuAD 2.0) is presented in detail in Section 4. Section 5 presents the systems and results proposed by participating teams. In Section 6, we provide further analysis of the challenge results. Finally, Section 7 summarizes the findings of the VLSP 2021-ViMRC Challenge and suggests future research directions.

2. Background and Related Works

Machine Reading Comprehension (MRC) has attracted many researchers in developing machine learning-based MRC models after the introduction of SQuAD (a large-scale and high-quality dataset) [5]. The growth in human-annotated datasets and computing capabilities are key factors behind the dramatic progress in the machine reading comprehension models. Particularly, many of datasets are constructed for evaluating the machine reading comprehension task including extractionbased MRC datasets (SQuAD [5], SQuAD 2.0 [6], TriviaQA [7], and NewsQA [8]), abstractive MRC dataset (NarrativeQA [9], RECAM [22]), multiple-choices datasets (RACE [10] and MCTest [4]), and conversational reading comprehension dataset (CoQA [11] and ViCoQA [23]). In addition to the creation of the MRC datasets, various neural network techniques [12, 14, 24, 25] have been presented and made significant

progress in this field. Table 1 shows the comparison of different MRC datasets.

Various efforts to create Vietnamese MRC datasets have been conducted. UIT-ViQuAD [16], UIT-ViNewsQA [21] are two corpora for the extraction-based machine reading comprehension task in Vietnamese language. Besides, two Vietnamese QA systems [26, 27] were developed with automatic reading comprehension techniques. In addition, ViMMRC [28] and ViCoQA [29] are two Vietnamese corpora for multiple-choices reading comprehension and conversational reading comprehension, respectively. Besides, a few MRC and QA methods have been studied on Vietnamese MRC datasets, such as BERT [23], ViReader [15], XLMRQA [26], and ViQAS [27].

Passage: Mã máy nhị phân (khác với mã hợp ngữ) có thể được xem như là phương thức biểu diễn thấp nhất của một chương trình đã biên dịch hay hợp dịch, hay là ngôn ngữ lập trình nguyên thủy phụ thuộc vào phần cứng (ngôn ngữ lập trình thế hệ đầu tiên). Mặc dù chúng ta hoàn toàn có thể viết chương trình trực tiếp bằng mã nhị phân, việc này rất khó khăn và dễ gây ra những lỗi nghiêm trọng vì ta cần phải quản lý từng bit đơn lẻ và tính toán các địa chỉ và hằng số học một cách thủ công. Do đó, ngoại trừ những thao tác cần tối ưu và gỡ lỗi chuyên biệt, chúng ta rất hiếm khi làm điều này. *(English: Binary machine code (as opposed to assembly code) can be thought of as the most basic representation of a compiler or assembled program or as a hardware-dependent primitive programming language (the first generation programming). Although it is capable of building programs directly in binary, doing so would be complex and prone to major errors because we must handle every bit as well as compute addresses and constants. As a result, except for procedures requiring optimization and specialized debugging, we very rarely do this.)*	
Question 1	Dù có thể sử dụng mã máy nhị phân để lập trình, nhưng tại sao các lập trình viên lại không sử dụng nó? *(Why don't programmers utilize binary machine code, even though it is possible?)*
Answer	những thao tác cần tối ưu và gỡ lỗi chuyên biệt *(except for procedures requiring optimization and except for procedures requiring optimization and specialized debugging)*
Answer start	493
Question 2	Ngôn ngữ lập trình thế hệ đầu tiên là ngôn ngữ gì? *(What is a first-generation programming language?)*
Answer	Mã máy nhị phân *(Binary machine code)*
Answer start	0
Question 3	Ngôn ngữ lập trình hợp ngữ đầu tiên là ngôn ngữ gì? *(What is the first assembly language?)*
Answer	-
Answer start	-
Plausible answer	Mã máy nhị phân *(Binary machine code)*
Plausible answer start	0

Figure 1. Several passage-question-answer triples extracted from the dataset.

Ultimately, SQuAD 2.0 [6], and NewsQA [8] are two corpora claiming the challenge of unanswerable questions in machine reading comprehension tasks, which are similar to our challenge. In general, extraction-based MRC requires computer understanding and retrieving the correct answer from the reading texts, which can evaluate the comprehension of the computer's natural language texts. However, the computer not only answers given questions as usual but also knows which questions are unanswerable. Our purpose in the challenge is to construct a dataset to evaluate the ability of the computer on both answerable and unanswerable questions for the extraction-based machine reading comprehension task.

Table 1. Benchmark of existing reading comprehension datasets, including UIT-ViQuAD

Dataset	Language	Size	Answerable	Unaswerable
SQuAD1.1 [5]	English	100k+	✓	
SQuAD2.0 [6]	English	150k+	✓	✓
KorQuAD [17]	Korean	70k+	✓	
SberQuAD [18]	Russian	50k+	✓	
CMRC-2018 [1]	Chinese	20k+	✓	
FQuAD1.1 [19]	French	60k+	✓	
FQuAD2.0 [20]	French	60k+	✓	✓
UIT-ViNewsQA [21]	Vietnamese	23k+	✓	
UIT-ViQuAD 1.0 [16]	Vietnamese	22k+	✓	
UIT-ViQuAD 2.0 (Ours)	Vietnamese	35k+	✓	✓

3. The VLSP 2021-ViMRC Challenge

3.1. Task Definition

This task aims to enable the ability of computers to understand natural language texts and answer relevant questions from users. The task is defined as below:

• Input: Given a text $T = \{t_1, \dots, t_n\}$ and a question $Q = \{q_1, \dots, q_m\}$ which can be answerable or unanswerable.

• Output: An answer $A = [a_s, a_e]$, where $0 \leq a_s \leq a_e \leq n$ can be a span $t_{a_s}, \dots, t_{a_e}$ extracted directly from *T* or empty if no answer is found.

The answers returned by the system are represented as answer spans by character level and are extracted from the reading text. The spans begin with an index indicating the location of the answer in the reading text. The end of the spans is an index determined by the sum of the start index and the length of the answers text. Nevertheless, the question in this task consists of answerable and unanswerable questions (as described in Figure 1), which is more difficult than the ViQuAD dataset [16]. According to Figure 1, the first and the second questions are answerable questions. The answers are directly extracted from the reading passage (highlighted by colors in the reading passage. The blue one is the answer for the first question, and the red one is the answer for the second question). The third question is unanswerable, however, according to Rajpurkar et al. [6], the plausible answers are added to the dataset to make it more diverse and create the challenge for current machine reading comprehension to enhance the ability of computers for understanding natural languages.

3.2. Evaluation Metrics

Following the evaluation metrics on SQuAD2.0 [6], we use EM and F1-score as evaluation metrics for Vietnamese machine reading comprehension. These evaluation metrics are described as below:

• Exact Match (EM): If the characters of the MRC system's predicted answer exactly match the characters of (one of) the gold standard answer(s), EM = 1 for each question-answer pair; otherwise, EM = 0. The EM metric is a strict all-or-nothing measurement, with a score of 0 for a single character error. If the method predicts any textual span as an answer

when evaluating against an unanswerable question, the question receives a zero score.

• F1-score: F1-score is a popular metric for natural language processing and is also used in machine reading comprehension. F1-score estimated over the individual tokens in the predicted answer against those in the gold standard answers. The F1-score is based on the number of matched tokens between the predicted and gold standard answers.

The final ranking is evaluated on the private test set, according to the F1-score (EM as a secondary metric when there is a tie).

3.3. Schedule and Overview Summary

Table 2 shows important dates of the VLSP 2021-ViMRC Challenge. It lasted for two months, during which the participating teams spent 27 days developing the models.

Table 2. Schedule of the VLSP 2021-ViMRC Challenge

Time	Phase
October 1st	Trial Data
October 5th	Public test
October 25th	Private test
October 27th	Competition end
November 15th	Submission deadline
December 15th	Notification of acceptance
December 28th	Camera-ready due

Besides, Table 3 describes an overview of participants who joined the competition. To get access to the system, each team must nominate a delegate, and register with the organizers. Only delegates of teams can submit the result to the system (as shown on the leaderboard).

Table 3. Participation summary of the VLSP 2021 ViMRC Challenge

Metric	Value
#Registration Teams	77
#Joined Teams	42
#Signed Data Agreements	42
#Submitted Teams	24
#Paper Submissions	6

Finally, Table 4 shows the statistical information about the results of participants by F1 and EM scores. Overall, the highest EM score is not higher than 78 percent, while the highest F1 score is 84.24 percent. Both the highest F1 and EM scores come from the public test. However, the results on the private test set are lower. Notably, the standard deviation of results by F1 and EM scores on the private test set is significantly higher than the public test set, which means the results between participating teams are different.

Table 4. Results overview of the VLSP 2021-ViMRC Challenge

	Public Test	Private Test	Overall
Total Entries	551	39	590
Highest F1	84.24	77.24	84.24
Highest EM	77.99	67.43	77.99
Mean F1	70.70	60.96	66.37
Mean EM	61.13	50.47	56.39
Std. F1	12.34	23.38	18.52
Std. EM	12.57	20.82	17.38

4. Dataset Construction

We proposed a new dataset named UITViQuAD 2.0 for this task, the latest version of the Vietnamese Question Answering Dataset. This dataset includes questions from the first version of UIT-ViQuAD [16] and nearly 13,000 newly human-generated questions which are unanswerable (see Section 4.1) and answerable (see Section 4.2). Instead of generating unanswerable questions from scratch like SQuAD 2.0 [6], we transform answerable questions into unanswerable questions. We randomly sample one-half of answerable questions in the original dataset and ask our annotators to transform these questions into unanswerable ones, which are impossible to answer given the information of the passage. The answers for answerable questions are then used as the plausible answers for unanswerable questions. This ensures that the unanswerable questions are similar to answerable ones, and the quality of plausible answers for unanswerable questions is high enough for further research into the behavior of Question Answering models.

4.1. Generating Unanswerable Questions

To generate unanswerable questions, we do a strict process of two phases: (1) unanswerable

question creation and (2) unanswerable question validation.

4.1.1. Unanswerable Question Creation

We hire 13 high-quality annotators for the process of generating unanswerable questions, most of whom have experience in annotating different datasets in Vietnamese Natural Language Processing. Our hired annotators are carefully trained in 6 phases in 10 days with 30 questions each phase. In the first 2 phases, we mainly focus on getting our annotators familiar with the task. In the next 4 phases, annotators are asked to create questions with a diverse range of unanswerable categories.

We did this by having our 13 annotators transform the same set of questions. Then, when more than two annotators have the same way of transforming an answerable question into an unanswerable one, these annotators will be asked to transform that question again. The result of this process is that there are many categories of unanswerable questions in our dataset, such as Antonym, Overstatement, Understatement, Entity Swap, Normal Word Swap, Adverbial Clause Swap, Modifiers Swap. This proposes new challenges to Vietnamese Machine Reading Comprehension researchers. Table 5 presents categories of unanswerable questions in UIT-ViQuAD 2.0.

We include all answerable questions, besides newly generated unanswerable ones, from the previous version of our dataset. This gives us a dataset with the proportion of roughly one unanswerable question per 2 answerable questions. Table 6 summarizes the dataset's overall statistics.

4.1.2. Unanswerable Question Validation

Before publishing the dataset for the evaluation campaign, we have carefully validated newly unanswerable questions following the procedure inspired by Nguyen et al. [16]. To help annotators gradually be better at generating new unanswerable questions, after generating every 3,000 unanswerable questions, we asked our annotators to self-validate the questions that they have generated before and write short documents to reflect on their errors. This effort minimizes the possibility that our annotators repeat their errors too many times.

To further reduce the error rate in our unanswerable questions, we have a separate phase of cross-validating after finishing creating 12,000 unanswerable questions. We hired ten annotators who had generated over 1,000 unanswerable questions during the phase of generating new samples for this phase. This effort helped filter out the annotators who have little experience in annotating unanswerable questions to reduce the noise during the validation phase. Our team then investigated and confirmed every error detected by annotators. To maximize the probability of detecting errors in newly generated unanswerable questions, we provide our annotators with incentives to carefully check for the errors in the dataset as we additionally reward them on each error they correctly detect.

4.2. Additional Difficult Answerable Questions

In addition to answerable questions from UIT-ViQuAD 1.0, we also hire five annotators, who have experiences in doing researches with Vietnamese natural language processing and clearly understand different reasoning skills [30] that is important to evaluate the comprehension ability of models to annotate more challenging answerable questions, which requires models more reasoning ability to correctly answer. The selected annotators are then encouraged to spend at least 3 minutes per question. When generating this set of questions, our purpose is to propose more challenges to researchers in the VLSP 2021 Evaluation Campaign and encourage further analysis on the effects of unanswerable questions in future works.

4.3. Overview Statistics of UIT-ViQuAD 2.0

The general statistics of the datasets are given in Table 6. UIT-ViQuAD 2.0 comprises 35,990 question-answer-passage triples (including 9,217 unanswerable questions). The organizers provide training, public test, and private test sets for the participating teams. For public and

private test sets, we only provide passages and their questions without answers to the teams.

5. Systems and Results

5.1. Baseline System

Following Devlin et al. [13], we adopt transfer learning based on BERT (Bidirectional Encoder Representations from Transformers) for our baseline system. To adapt to our dataset, we slightly modify the run squad.py script while keeping the majority of the original code. mBERT is trained on 104 languages, including Vietnamese. In addition, we use the transformers library by Hugging Face to fine-tune mBERT for our question-answering dataset. We fine-tuned the parameters to suit our dataset in the training process as well as the model evaluation process. For the baseline system, we used an initial learning_rate of 3e-5 with a batch_size of 32 and trained for two epochs. The max_seq_length and doc_stride are set to 384 and 128.

5.2. Challenge Submissions

The AIHUB platform (https://aihub.vn/) was used to manage all submissions. We received entries from 24 teams for the public test, while for the private test, we received submissions from 18 teams. The systems using the pre-trained language model XLM-R achieve SOTA results. Six of these teams had their system description papers submitted. Each of them is briefly described below.

5.2.1. The Vc-tus Team

With addressing unanswerable questions, the Vc-tus team presents a novel Vietnamese reading comprehension system based on Retrospective Reader [31]. Furthermore, they concentrate on improving answer extraction ability by utilizing attention mechanisms efficiently and boosting representation capacity through semantic information processing. They also offer an ensemble strategy for achieving significant improvements in single model results. Their method won the first place in the VLSP 2021 – ViMRC Challenge.

5.2.2. The ebisu_uit Team

The ebisu_uit team presents a novel method for training Vietnamese reading comprehension. To tackle the Machine reading comprehension task in Vietnamese, they apply BLANC (BLock AttentioN for Context prediction) [32] on pre-trained language models. With this strategy, this model produced good results. This approach achieved 77.22 percent of F1-score on the private test with the MRC task at the VLSP 2021 – ViMRC Challenge, placing the second rank overall.

5.2.3. The F-NLP Team

To learn the correlation between a start answer index and an end answer index in pure-MRC output prediction, the F-NLP team presents two types of joint models for answerability prediction and pure-MRC prediction with/without a dependence mechanism. They also use ensemble models and a verification approach that involves choosing the best answer from among the top K answers offered by different models.

5.2.4. The UIT-MegaPikachu Team

The UIT-MegaPikachu team proposes a new system which employs simple yet highly effective method. The system uses a strong pre-trained language model (PrLM) XLMRoBERTa [14], combined with filtering results from multiple outputs to produce the final result. This system generated around 5-7 outputs and chose the answer with the highest number of repetitions as the final predicted answer.

Table 5. Categories of unanswerable questions in UIT-ViQuAD 2.0

Reasoning	Description	Example
Antonym	Antonym used	**Sentence**: Vào năm 1171, Richard khởi hành đến Aquitaine với mẹ mình và Henry phong ông là Công tước xứ Aquitaine theo yêu cầu của Eleanor. (*In 1171, Richard departed for Aquitaine with his mother, and Henry made him Duke of Aquitaine at Eleanor's request*) **Original question**: Richard khởi hành đến Aquitaine với mẹ vào năm nào? (*In what year did Richard depart for Aquitaine with his mother?*) Unanswerable question: Richard khởi hành từ Aquitaine với mẹ vào năm nào? (*In what year did Richard depart from Aquitaine with his mother?*)
Overstatement	Word that has similar meaning but with a higher shades of meaning is used	**Sentence**: Ngày 9 tháng 11 năm 1989, vài đoạn của Bức tường Berlin bị phá vỡ, lần đầu tiên hàng ngàn người Đông Đức vượt qua chạy vào Tây Berlin và Tây Đức. (*On November 9, 1989, several parts of the Berlin Wall were collapsed, and for the first time thousands of East Germans crossed into West Berlin and West Germany.)* **Original question**: Bức tường Berlin đã bị sụp đổ một vài đoạn vào ngày nào? (*On which date were some parts of Berlin Wall collapsed?*) **Unanswerable question**: Bức tường Berlin đã bị sụp đổhoàn toàn vào ngày nào? (*On which date was Berlin Wall completely collapsed?*)
Understatement	Word that has similar meaning but with a lower shades of meaning is used	**Sentence**: Quân đội Nhật Bản chiếm đóng Quảng Châu từ năm 1938 đến 1945 trong chiến tranh thế giới thứ hai. (*The Japanese army captured Guangzhou from 1938 to 1945 during the second world war.*) **Original question**: Khi Chiến tranh Thế giới thứ hai xảy ra thì Quảng Châu bị nước nào chiếm đóng? (*During the World War II, Guanzong was captured by which country?*) **Unanswerable question**: Khi Chiến tranh Thế giới thứ hai xảy ra thì Quảng Châu bị nước nào đe dọa? (*During the World War II, Guanzong was attacked by which country?*)
Entity Swap	Entity replaced by other entity	**Sentence**: Là cảng Trung Quốc duy nhất có thể tiếp cận được với hầu hết các thương nhân nước ngoài, thành phố này đã rơi vào tay người Anh trong chiến tranh nha phiến lần thứ nhất. (*As the only Chinese port accessible to most foreign merchants, the city fell to the British during the First Opium War.*) **Original question**: Trong cuộc chiến nào thì Anh Quốc đã chiếm được Quảng Châu? (*In which war did Britain capture Guangzhou?*) **Unanswerable quetion**: Trong cuộc chiến nào thì Nhật đã chiếm được Quảng Châu? (*In which war did Japan capture Guangzhou?*)
Normal Word Swap	A normal word replaced by another normal word	**Sentence**: Sự phát hiện của Hofmeister năm 1851 về các thay đổi xảy ra trong túi phôi của thực vật có hoa [...] (*Hofmeister's discovery in 1851 of changes occurring in the embryo sac of flowering plants [...]*) **Original question**: Năm 1851 nhà sinh học Hofmeister đã tìm ra điều gì ở thực vật có hoa? (*In 1851, the biologist Hofmeister discovered what in flowering plants?*) **Unanswerable question**: Năm 1851 nhà sinh học Hofmeister đã công nhận điều gì ở thực vật có hoa? (*In 1851, the biologist Hofmeister accepted what in flowering plants?*)
Adverbial Clause Swap	Adverbial clause replaced by another adverbial clause related to the context	**Sentence**: Trước đó Phạm Văn Đồng từng giữ chức vụ Thủ tướng Chính phủ Việt Nam Dân chủ Cộng hòa từ năm 1955 đến năm 1976. Ông là vị Thủ tướng Việt Nam tại vị lâu nhất (1955–1987). Ông là học trò, cộng sự của Chủ tịch Hồ Chí Minh. (*Pham Van Dong previously held the position of Prime Minister of the Democratic Republic of Vietnam from 1955 to 1976. He was the longest-serving Prime Minister of Vietnam (1955-1987). He was a student and collaborator of President Ho Chi Minh.*) **Original question**: Giai đoạn năm 1955-1976, Phạm Văn Đồng nắm giữ chức vụ gì? (*In the period 1955-1976, what position did Pham Van Dong hold?*) **Unanswerable question**: Khi là cộng sự của chủ tịch Hồ Chí Minh, Phạm Văn Đồng nắm giữ chức vụ gì? (*As a collaborator of President Ho Chi Minh, what position did Pham Van Dong hold?*)
Modifiers Swap	Modifier of one word in the given context is used for another word	**Sentence**: Các phần mềm giáo dục đầu tiên trong lĩnh vực giáo dục đại học (cao đẳng) và tập trung được thiết kế chạy trên máy tính đơn (hoặc các thiết bị cầm tay). Lịch sử của các phần mềm này được tóm tắt trong SCORM 2004 2nd edition Overview (phần 1.3) (*The first educational software in the field of higher education (college) and concentration was designed to run on a single computer (or portable devices). The history of these software is summarized in SCORM 2004 2nd edition Overview (section 1.3).*) **Original question**: Lịch sử của các phần mềm giáo dục đầu tiên trong lĩnh vực giáo dục đại học (cao đẳng) được tóm tắt, ghi nhận ở đâu? (*Where did the history of the first educational software in the field of higher education (college) was summarized and recorded?*) **Unanswerable quetion**: Lịch sử của các phần mềm giáo dục trong lĩnh vực giáo dục đại học (cao đẳng) được tóm tắt, ghi nhận đầu tiên ở đâu? (*Where did the history of the educational software in the field of higher education (college) was first summarized and recorded?*)

Table 6. Overview Statistics of UIT-ViQuAD 2.0

	Train	Public Test	Private Test	All
#Articles	138	19	19	176
#Passages	4,101	557	515	5,173
#Total questions	28,457	3,821	3,712	35,990
#Unanswerable questions	9,217	1,168	1,116	11,501
Average passage length	179.0	167.6	177.3	177.6
Average answerable question length	14.6	14.3	14.7	14.6
Average unanswerable question length	14.7	14.0	14.5	14.6

5.2.5. The UITSunWind Team

The UITSunWind team introduces a new approach to solve the task at the VLSP 2021 – ViMRC Challenge. A novel system MRC4MRC using XLM-RoBERTa includes two main components. On the public-test set, the MRC4MRC based on the XLM-RoBERTa pre-trained language model achieves 79.13 percent in F1-score and 69.72 percent in Exact Match. Despite being among the top 5 models, the EM-based performance on answerable questions is the best on the private test. The XLM-RoBERTa language model outperforms the strong PhoBERT language model in their experiments.

5.2.6. The HN-BERT Team

The HN-BERT team offers an unsupervised passage selector that shortens a given passage while retaining answers in related passages.

In the corpus of the VLSP 2021 – ViMRC Challenge, they also applied a variety of experimental techniques, such as unanswerable question sample selection and different adversarial training methodologies, which enhanced performance by 2.5 percent in EM and 1 percent in F1-score.

5.3. Human Performance

To estimate the human performance of this task, we employ a team to answer a data set of 100 samples from the public test set and 100 samples from the private test set. There are four annotators, and two of them work on each data set doing the same thing.

In each instance, we have a passage with a question. The annotator must answer the question using the information in the passage. If there is no answer, it means the question is unanswerable, and then mark "true" in the field "is unanswerable". Following the answering phase, we compute the human accuracy by F1-score and exact match scores for both public and private tests.

To calculate human performance, we use the method given in SQuAD2.0 [6]. We have four responses per question in the ground truth. Thus we choose the final ground truth by majority voting and prefer the shortest answer to be the last ground truth, as explained in SQuAD2.0. After obtaining the gold response, we compute the F1 and EM scores in pairs of human-answering and gold answers with the two annotators who previously answered on the public test set. Then, by averaging the results of the two annotators, we compute the final F1 and EM scores of human performance on the public test. The computation is carried out on the private test in the same manner. As a result, the final F1 and EM scores of human performance are 87.34% and 82.85% on the public test set, respectively, and 81.82% and 75.50% on the private test set.

5.4. Experimental Results

According to our statistics, a total of 24 teams registered to submit their results. These teams from prestigious universities, companies, and organizations participate in the Vietnamese Machine Reading Comprehension task of the VLSP 2021-ViMRC Challenge. And then, out of the 24 teams participating in the development phase of their system on the public test we selected 18 teams that excelled against the baseline to further evaluate their system in the private test. The results of the teams in the two rounds are aggregated and shown in Table 7. The ranking results of the team are based on F1

points for both rounds. In the public test round, our mBERT baseline model achieved 63.03% on F1 and 53.55% on EM. There were 18 teams with results that outperformed the results of the baseline according to F1. Overall, 14 teams with F1 scores above 70% and 5 teams with over 80%. Specifically, we found the top three teams in the public test, NLP_HUST, NTQ, and ebisu_uit, with F1 results of 84.24%, 84.09%, and 82.63%, respectively. It can be seen that the results of the two top teams in the rankings have very close results. The difference between these two teams is not more than 0.2%. Additionally, the NTQ team's model scored slightly lower in F1 than NLP_HUST, but their model achieved the highest EM performance of 77.99%.

Regarding the private test round, the baseline model's results achieved an F1 score of 60.34% and 49.35% on the EM score. Out of the 18 teams that passed the public test, 14 continued to participate in evaluating their system on the private test set. There have been many unexpected changes in the results of the teams' submissions, especially the way the top three teams appeared. While only placing in 5th with 81.01% of F1 on the public test set, team vc-tus took 1st position in the private test round with an F1 score of 77.24%. Besides, the ebisu_uit team maintains a stable level on the model they trained from the public test round to the private test round. They have kept 2nd place in the rankings with their F1 score of 77.22%. Once again, we can see that the results are not much different between the 1st and 2nd place teams.

Furthermore, ebisu_uit is also the team with the highest results on the EM measure with 67.43%. If we take a look at the F-NLP team, it shows a similar trend with vc-tus. Remaining 5th in the public test round, their system helped them finish this task at 3rd with 76.46% of F1 score. Generally, all the teams in this round were having trouble with the private test set since its difficulty had increased significantly. As a result, the submission results of the teams are reduced considerably compared to the public test round.

Table 7. Final results on the public and private test sets. Participating teams are ranked by their highestF1-score

Public Test Phase			Private Test Phase		
	F1	**EM**		**F1**	**EM**
Human	87.335	81.818	**Human**	82.849	75.500
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
UIT-MegaPikachu	78.637	68.804	914-clover	73.027	61.853
914-Clover	78.515	69.013	NTQ	72.863	60.938
Big Heroes	78.491	68.150	Hey VinMart	70.352	57.786
PhoKho-UIT	75.894	65.533	PhoKho-UIT	70.198	58.378
HN-BERT	75.842	63.544	HN-BERT	70.100	56.466
Hey VinMart	75.759	64.590	Deep-NLP	69.220	59.429
Deep-NLP	74.767	66.789	ABC	63.625	55.280
ABC	69.287	57.864	**BASELINE**	**60.338**	**49.353**
ct-nlp	68.971	58.859			
tpp	68.484	57.786			
S-NLP	67.589	65.140			
BASELINE	**63.031**	**53.546**			

6. Result Analysis

To gain a deeper insight into machine reading comprehension and question answering in Vietnamese, we analyze the results based on the 5 most powerful models at the VLSP 2021 ViMRC Challenge.

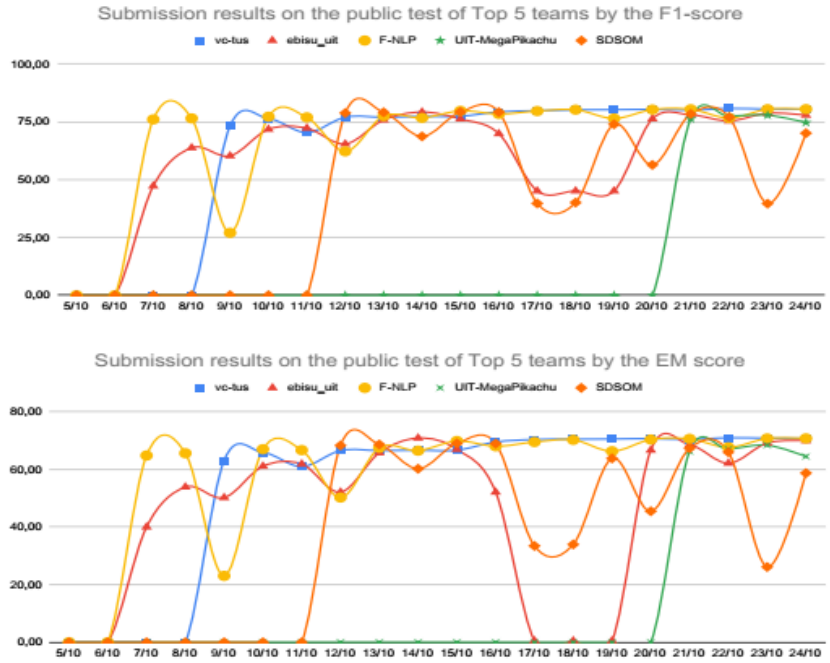

Figure 2. Submission progress of the Top 5 teams on the public test phase.

Table 8. Final results on answerable and unanswerable questions of the private test set

Teams	Models	Answerable		Unanswerable		Overall	
		EM	F1	EM	F1	EM	F1
vc-tus	Retrospective Reader + XLM-R (Ensemble)	57.67	73.54	85.84	85.84	66.14	**77.24**
ebisu_uit	BLANC + XLM-R/SemBERT (Ensemble)	56.59	70.59	**92.65**	**92.65**	**67.43**	77.22
F-NLP	XLM-R (Ensemble)	58.78	**75.66**	78.32	78.32	64.66	76.46
UIT-MegaPikachu	XLM-R (Single)	58.82	74.63	80.47	80.47	65.33	76.39
UITSunWind	XLM-R + BiLSTM (Ensemble)	**58.94**	74.26	78.67	78.67	64.87	75.59
HN-BERT	PhoBERT_Large+R3F+CS (Single)	47.50	66.99	77.33	77.33	56.47	70.10
Baseline	mBERT (Single)	41.72	57.43	67.11	67.11	49.35	60.34

6.1. Competition Progress Analysis

Figure 2 illustrates the submission progress of the top 5 teams on the public test from October 5, 2021, to October 24, 2021. In this phase, we allow 10 submissions per day. However, according to Figure 2, the submission results on both F1 and EM scores are not stable, which oscillates within the submission time. Besides, the results by EM score are no higher than 80%, indicating the challenge in the dataset for the participants.

In addition, Figure 3 illustrates the final submission results of participant teams. The private test started on October 25, 2021, and ended on October 27, 2021. Within 3 days of submission, the results on the F1 score do not change too much. Both F1 and EM scores achieved by participants are not higher than 80% in this phase. Especially, for the final results, the team name ebisu_uit has a lower result than the vc_tus team but achieved a higher result on the EM score. It can be seen from the chart that the team vc_tus achieved the best results by the F1-score, and the team ebisu_uit achieved the best result by the EM score, which placed the 1st and 2nd in the competition.

6.2. Answerable vs. Unanswerable Analysis

To better understand the ability of the MRC systems to answer questions, we analyze human performance and the experimental results of the

baseline model and the participating teams. Table 8 shows final results on answerable and unanswerable questions of the private test set, evaluated on EM and F1 scores. As seen from the table, performances on unanswerable questions are always higher than on answerable questions. The ebisu_uit team achieved the best performance on unanswerable questions with over 92% of F1. However, the F-NLP and UITSunWind teams achieved the highest scores on the answers with 75.66% of F1 and 58.82% of EM, respectively. Interestingly, the vc-test team did not obtain the best performance on unanswerable and answerable questions, but this team achieved the best performance on the overall F1-score because they balanced the performances between the two types of questions better than the other teams.

6.3. Challenging Question Examples

We select several typical examples of answerable and unanswerable questions that make it difficult for the models proposed by the participating teams. Figure 4 presents several examples and explanations that the models failed to predict correct answers. We will explore more complex questions inspired by the works [33, 34].

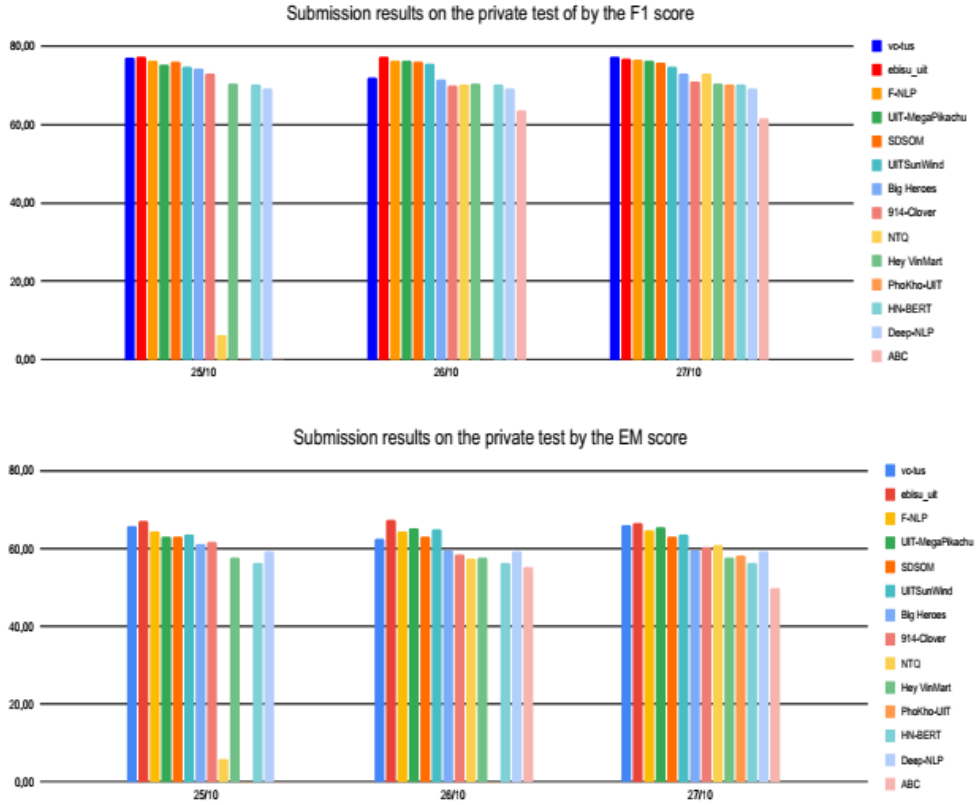

Figure 3. Submission progress of the teams who results are higher than baseline score on the private testphase.

Example	Explanation
Passage: Sau khi sinh ra, Edward được một nhũ mẫu có tên Mariota hoặc Mary Maunsel chăm sóc trong vài tháng trước khi bà ta phát bệnh, và Alice de Leygrave trở thành dưỡng mẫu của ông. Ông có thể hoàn toàn không biết mặt người mẹ ruột Eleanor đã ở Gascony với cha ông trong những năm đầu đời của ông. (*After his birth, Edward was cared for by a nanny named Mariota or Mary Maunsel for several months before she became ill, and Alice de Leygrave became his nanny. He may be completely unaware of his biological mother Eleanor who was in Gascony with his father during his early years.*) **Question**: Thân mẫu của Edward II là ai? (Who is Edward II's mother?) **Correct answer**: Eleanor	Questions require multiple reasonings, challenging all MRC systems in the shared task. For example, this question requires co-reference, lexical knowledge and external knowledge to find the correct answer. Co-reference: Ông (He) is linked to Edward. Lexical knowledge: Mẹ ruột (mother) is the same meaning of Thân mẫu. External knowledge: Edward in this context is Edward II, not Edward I, II, etc.
Passage: Tháng 6 năm 2010, Apple cho ra mắt chiếc iPhone 4, chiếc smartphone thiết kế cao cấp với hai mặt kính và khung kim loại, màn hình độ phân giải cao nhất với độ phân giải 960x640 pixel được gọi là màn hình Retina, cùng với vi xử lý Apple A4 (ARM Cortex A8) mạnh mẽ và bộ nhớ Ram 512 MB và camera nâng cấp lớn lên đến 5 MP quay phim 720p với 30 khung hình 1 giây và có đèn Led ở đằng sau, đây cũng là chiếc smartphone được trang bị camera trước với độ phân giải VGA và tính năng gọi video call lần đầu tiên có tên là Facetime độc quyền của Apple. (*In June 2010, Apple released the iPhone 4, a premium design smartphone with two glass sides and a metal frame, the highest resolution screen with a resolution of 960x640 pixels called Retina display, a powerful Apple A4 (ARM Cortex A8) processor and 512MB RAM, and a sizeable upgraded camera up to 5 MP with 720p video recording at 30 frames per second and with LED lights on the back, this is also a smartphone that is designed with a front camera with VGA.*) **Question**: Dung lượng bộ nhớ của Apple A4 là bao nhiêu? (How much memory does the Apple A4 have?) **Correct answer**: “” **Predicted answer by top 5 teams**: 512 MB	There are many entity objects in the context, the relationship between these objects is ambiguous. In this example, all MRC Systems fail to understand the relation ambiguity between 512MB with Apple A4 or RAM.

Figure 4. Several examples and explanations that the models failed to predict correct answers. Theexample texts in the ViQuAD 2.0 dataset are taken from the Vietnamese Wikipedia.

7. Conclusion and Future Works

The VLSP 2021-ViMRC Challenge on Machine Reading Comprehension for Vietnamese has been organized at the VLSP 2021. Despite the fact that 77 teams had signed up to get the training datasets, only 24 teams were able to submit their results. Because several teams enrolled for many challenges at the VLSP 2021, the other teams may not have enough time

to explore MRC models. This challenge provides valuable resources for developing Vietnamese machine reading comprehension, question answering, question generation (QG), and other AI applications using MRC, QA, and QG models.

To increase performance in machine reading comprehension systems, in the future, we intend to increase the amount and quality of annotated questions. In addition, we also make difficult questions based on findings proposed by the research works [35, 33, 36]. UIT-ViQuAD 2.0 can also be used to evaluate various other NLP tasks: question answering that uses retriever-reader techniques [37, 27], question generation [38], and information retrieval [39]. We will explore more complex questions inspired by the works [33, 34]. Finally, UIT-ViQuAD 2.0 will be provided to evaluate MRC, QA, and QG models, including the training set, the development set (public test set) and the test set (private test set).

Acknowledgments

The authors would like to thank the team of aihub.vn, and the annotators for their hard work to support the VLSP 2021-ViMRC Challenge. The VLSP Workshop was supported by organizations: VINIF, Aimsoft, Zalo, Bee, and INT2, and universities: VNUHCM University of Information Technology, VNU University of Science, and VNU University of Engineering and Technology. Kiet Van Nguyen was funded by Vingroup JSC and supported by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Institute of Big Data, code VINIF.2021.TS.026.

References

[1] 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, arXiv preprint arXiv:1810.07366.

[2] A. Fisch, A. Talmor, R. Jia, M. Seo, E. Choi, D. Chen, Mrqa 2019 Shared Task: Evaluating Generalization in Reading Comprehension, arXiv:1910.09753.

[3] B. Zheng, X. Yang, Y. P. Ruan, Z. Ling, Q. Liu, S. Wei, X. Zhu, Semeval 2021 Task 4: Reading Comprehension of Abstract Meaning, arXiv preprint arXiv:2105.14879.

[4] M. Richardson, C. J. Burges, E. Renshaw, MCTest: A Challenge Dataset for the OpenDomain Machine Comprehension of Text, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 193–203.

[5] 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, Association for Computational Linguistics, 2016, pp. 2383–2392.

[6] P. Rajpurkar, R. Jia, P. Liang, Know What You Don't Know: Unanswerable Questions for Squad, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 2, 2018, pp. 784–789.

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

[8] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, K. Suleman, Newsqa: A Machine Comprehension Dataset, in: Proceedings ofthe 2nd Workshop on Representation Learning for NLP, Association for Computational Linguistics, 2017, pp. 191–200.

[9] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, E. Grefenstette, The Narrative QA Reading Comprehension Challenge, Transactions of the Association for Computational Linguistics, Vol. 6, 2018, pp. 317–328.

[10] 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.

[11] S. Reddy, D. Chen, C. D. Manning, Coqa:A Conversational Question Answering Challenge, Transactions of the Association for Computational Linguistics, Vol. 7, 2019, pp. 249–266.

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

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pretraining 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.

[14] 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, 2016, arXiv:1911.02116.

[15] K. V. Nguyen, N. Duy 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, pp. 1–19.

[16] K. V. 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, International Committee on Computational Linguistics, 2020, pp. 2595–2605.

[17] S. Lim, M. Kim, J. Lee, Korquad 1. 0: Korean QA Dataset for Machine Reading Comprehension, arXiv:1909.07005.

[18] P. Braslavski, Sberquad Russian Reading Comprehension Dataset: Description and Analysis, in: Experimental IR Meets Multilinguality, Multimodality and Interaction: 11th International Conference of the CLEF Association, Proceedings, Springer Nature, Vol. 12260, 2020, p. 3.

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

[20] Q. Heinrich, G. Viaud, W. Belblidia, Fquad2. 0: French Question Answering and Knowing That You Know Nothing, arXiv:2109.13209.

[21] 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 preprint arXiv:2006.11138.

[22] B. Zheng, X. Yang, Y. P. Ruan, Z. Ling, Q. Liu, S. Wei, X. Zhu, SemEval-2021 task 4: Reading Comprehension of Abstract Meaning, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, 2021, pp.37–50.

[23] S. T. Luu, K. V. Nguyen, A. G. T. Nguyen, N. 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), 2021, pp. 282–287.

[24] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pretraining of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805.

[25] N. V. Tu, L. A. Cuong, A Deep Learning Model of Multiple Knowledge Sources Integration for Community Question Answering, , VNU Journal of Science: Computer Science and Communication Engineering, 37(1), 2021.

[26] K. V. Nguyen, N. Duy Nguyen, P. N. T. Do, T. V. Huynh, A. G. T Nguyen, N. L. T. Nguyen,Xlmrqa: Open-domain question answering on vietnamese wikipedia-based textual knowledge source, arXiv:2204.07002.

[27] 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. International Journal of Machine Learning and Cybernetics.

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

[29] S. T. Luu, M. N. Bui, L. D. Nguyen, K. V. Tran, K. Van Nguyen, N. L. T. Nguyen, Conversational Machine Reading Comprehension for Vietnamese Healthcare Texts, in: K. Wojtkiewicz, J. Treur, E. Pimenidis, M. Maleszka (Eds.), Advances in Computational Collective Intelligence, Springer International Publishing, Cham, 2021, pp. 546–558.

[30] S. Sugawara, Y. Kido, H. Yokono, A. Aizawa, Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2017, pp.806–817.

[31] Z. Zhang, J. Yang, H. Zhao, Retrospective Reader for Machine Reading Comprehension, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 14506–14514.

[32] 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.

[33] S. Sugawara, K. Inui, S. Sekine, A. Aizawa, What

Makes Reading Comprehension Questions Easier? in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,2018, pp. 4208–4219.

[34] S. Sugawara, P. Stenetorp, K. Inui, A. Aizawa, Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8918–8927.

[35] S. Sugawara, Y. Kido, H. Yokono, A. Aizawa, Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2017, pp. 806–817.

[36] S. Sugawara, P. Stenetorp, A. Aizawa, Benchmarking machine reading comprehension: A psychological perspective, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 1, 2021, pp. 1592–1612.

[37] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading Wikipedia to Answer Open-Domain Questions, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2017, pp. 1870–1879.

[38] X. Du, J. Shao, C. Cardie, Learning to Ask: NeuRal Question Generation for Reading ComprehenSion, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2017, pp. 1342–1352.

[39] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense Passage Retrieval for Open-Domain Question Answering, arXiv:2004.04906.