Machine Reading Comprehension for Vietnamese Customer Reviews: Task, Corpus and Baseline Models

Tinh Pham Phuc Do, Ngoc Dinh Duy Cao, Nhan Thanh Nguyen, Tin Van Huynh, Kiet Van Nguyen
Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{20522020, 20521661, 20521701}@gm.uit.edu.vn and {tinhv,kietnv}@uit.edu.vn

Abstract

Customers spend much time researching product information before making a purchase. This problem can be partially addressed through Machine Reading Comprehension (MRC) on customer reviews. Nonetheless, to implement MRC effectively, benchmark corpora specific to the review domain in Vietnamese are lacking. Therefore, we proposed **ViRe4MRC**, the first benchmark corpus for evaluating review-based machine reading comprehension on customer reviews in Vietnamese. This corpus comprises 6,603 human-generated question-answer pairs from 2,174 customer reviews on smartphone and restaurant domains. We also evaluate the experimental results of monolingual language models: ViBERT, PhoBERT, and vELECTRA; multilingual language models: mBERT and XLM-RoBERTa (XLM-R). As a result, the XLM-R$_{Large}$ model, as the best model, achieved 44.25% Exact Match (EM) and 78.13% F1. Our corpus[1] is available for research purposes.

1 Introduction

Customer review data is challenging for machine reading comprehension (MRC) tasks. The linguistic features of the data contribute to the challenges related to understanding and processing information. Several features that cause difficulties include:

- Reviews may include grammatically incorrect sentences, slang, spelling errors, incorrect punctuation, and ambiguous evaluations. These factors require language-understand annotators, posing challenges in comprehending the review content for generating question-answer pairs. For example, the phrase **"giao hàg nhah"** (fast delivery) is a spelling mistake, which should be **"giao hàng nhanh"** (fast delivery).

- In the review domain, textual contexts frequently contain diverse information about products, services, or specific aspects. This presents challenges in processing and understanding multiple contexts to extract accurate answers.

- The Vietnamese language possesses linguistic characteristics such as diverse expressions, various sentence structures, and relatively complex vocabulary. These attributes add to the heightened complexity of comprehending and processing questions for the MRC task.

In MRC tasks, having a sufficiently large and diverse training corpus is important. Recently, there has been the appearance of machine reading comprehension corpora for Vietnamese UIT-ViNewsQA (Van Nguyen et al., 2022), ViMMRC (Nguyen et al., 2020b), WikiQA (Do et al., 2021), ViCoQA (Luu et al., 2021), UIT-ViQuAD 2.0 (Van Kiet et al., 2022), ViCoV19QA (Thai et al., 2022), ViQA-COVID (Anonymous, 2021), ViMQA (Le et al., 2022). However, there is not yet any MRC corpus on Vietnamese customer reviews. Thus, we create a novel MRC corpus for Vietnamese customer reviews.

In this paper, we created the **ViRe4MRC** corpus in the review-based MRC task, comprising 6,603 question-answer pairs with non-standard text genres in the smartphone and restaurant review domains. We hope to contribute a MRC corpus on customer review domains to the research community to develop new MRC tasks. We also hope that when our

[1] https://github.com/DoPhamPhucTinh/ViRe4MRC

study applies in the real world, it can support businesses to increase revenue and customer experience.

2 Related Works

Machine reading comprehension (MRC) appeared early. Lehnert (1977) created a question-answering program, QUALM. Hirschman et al. (1999) proposed an MRC system with a corpus containing 60 development and 60 test stories. In the past decade, MRC has developed strongly, first with the appearance of the MCTest corpus (Richardson et al., 2013) and then a series of supervised corpora and methods for this task.

There are many MRC corpus (Rajpurkar et al., 2016; Cui et al., 2019; Kočiský et al., 2018; Nguyen et al., 2020a). Based on the answer type, the MRC task is divided into four types: span prediction, free-form answer, cloze style and multiple choice (according to Chen (2018)). The number of corpora devoted to four types of question-answering tasks is increasing. However, most of the corpora are done in English or Chinese. Here are several well-known corpora for the question-answering task.

- MRC with span prediction: The number of corpora of this type has recently increased significantly, and many quality corpora have been published: NewsQA (Trischler et al., 2017) is a corpus on the domain of news articles, including more than 119,633 question-answer pairs from CNN news articles. SQuAD (Rajpurkar et al., 2016) comprises passages and question-answer pairs in Wikipedia. CMRC2018 (Cui et al., 2019), the corpus comprises 18,567 questions in Chinese on Wikipedia paragraphs. ReviewRC (Xu et al., 2019) comprises 2,596 questions from 959 reviews on laptop and restaurant domains.

- MRC with the free-form answer: This field is being researched and promoted. The output is an answer based on the content of the passage, not a string in the passage. There are corpora such as SearchQA (Dunn et al., 2017), the corpus comprises 140,196 question-answer pairs; NarrativeQA (Kočiský et al., 2018), the corpus comprises 46,765 question–answer pairs; DuReader (He et al., 2018), is a Chinese corpus, which comprises over 200,000 questions, nearly 420,000 answers and 1,000,000 documents.

- MRC with cloze style: several corpora for this type is the CNN/DAILY corpus (Hermann et al., 2015) comprises over 387,000 samples from the CNN website[2] and over 997,000 samples from the DAILY website[3]. The CBT corpus (Hill et al., 2016) comprises over 687,000 questions taken from books within the Gutenberg project[4].

- MRC with multiple-choice answers: RACE (Lai et al., 2017) comprises nearly 28,000 passages and 100,000 questions; COSMOS QA (Huang et al., 2019) comprises 35,588 questions and 21,866 passages. ReClor (Yu et al., 2020) comprises 6,138 questions.

For Vietnamese, the MRC task has also recently been interesting. Vietnamese is a language with fewer corpora compared to other languages, such as English and Chinese. UIT-ViQuAD (Nguyen et al., 2020a) is the first Vietnamese corpus for supervised learning-based machine reading comprehension and question answering, which motivated the development of various Vietnamese corpora are UIT-ViNewsQA (Van Nguyen et al., 2022), ViMMRC (Luu et al., 2023), ViWikiQA (Do et al., 2021), ViCoQA (Luu et al., 2021), UIT-ViQuAD 2.0 (Van Kiet et al., 2022), ViCoV19QA (Thai et al., 2022), ViQA-COVID(Anonymous, 2021), ViMQA(Le et al., 2022). The details of the corpora are presented in Table 1.

3 Corpus Creation

3.1 Corpus Creation Process

This section provides a more detailed description of the ViRe4MRC corpus creation process, illustrated in Figure 1. We collected reviews from the ViSD4SA (Thanh et al., 2021) corpus and streetcodevn.com website[6]. The following

[2] https://edition.cnn.com/
[3] https://www.dailymail.co.uk/home/index.html
[4] https://www.gutenberg.org/
[6] https://streetcodevn.com/blog/dataset

Corpus	Source	Domain	Number of questions	Type of corpus
ViMMRC (Nguyen et al., 2020b)	Vietnamese subject of students documents.	Education	2,783	Multiple-choice
UIT-ViQuAD (Nguyen et al., 2020a)	Vietnamese wikipedia articles	Open domain	23,074	Span prediction
UIT-ViCoQA (Luu et al., 2021)	VnExpress newspaper articles	Health	10,000	Span prediction
UIT-WikiQA (Do et al., 2021)	Vietnamese wikipedia articles	Open domain	23,074	Span prediction
ViQA-COVID (Anonymous, 2021)	CDC case reports	COVID-19	6,444	Span prediction (Multi span)
ViMQA (Le et al., 2022)	Vietnamese Wikipedia	Open domain	10,047	Span prediction (Multi hop)
UIT-ViNewsQA (Van Nguyen et al., 2022)	VnExpress newspaper articles	Health	22,057	Span prediction
ViRe4MRC (Ours)	ViSD4SA (Thanh et al., 2021) and streetcodevn.com[5]	Customer Reviews	6,603	span prediction

Table 1: Vietnamese corpora for machine reading comprehension, question answering, and question generation.

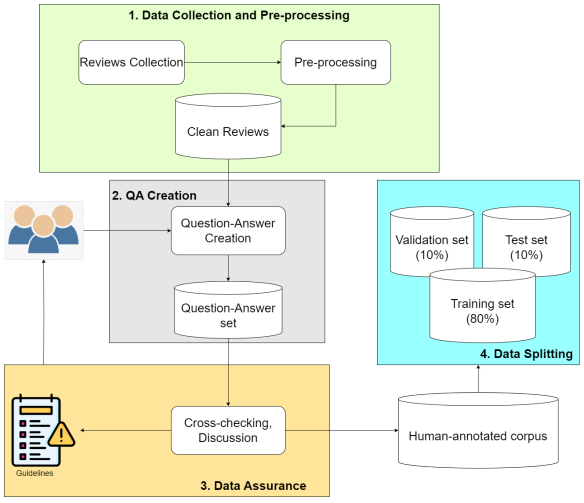

Figure 1: The corpus creation process.

phases are based on the creation process of two corpora, UIT-ViQuAD(Nguyen et al., 2020a) and UIT-ViNewsQA(Van Nguyen et al., 2022).

Three annotators generate question-answer pairs in ViRe4MRC to create an objective and natural corpus. All three annotators are students currently studying at the same university. Initially, each annotator familiarized themselves with the guidelines and accompanying examples. We then selected 50 reviews and assigned each annotator to independently generate question-answer pairs for these reviews. Each review required a minimum of 3 question-answer pairs. We review the question-answer pair they have created, identify errors, engage in discussions, and update the guidelines. We repeat this process multiple times until the question-answer pairs created by the annotators do not have any errors, following the provided guidelines. After completing the training phase, we officially create question-answer pairs for the corpus.

- Each annotator creates 200 question-answer pairs within 4-5 days. In the subsequent days, annotators cross-checked their results in pairs. During cross-checking, each person reviewed the results of the other person to ensure compliance with the guidelines and provided accurate feedback on any identified mistakes. In cases of disagreement or uncertainty, discussions

Sample 1

Review: Pizza khá lạ và khác so với kiểu truyền thống nên mình k thích lắm. **Mỳ ý sốt bò băm và sốt kem rất ngon**. Giá đồ ăn đúng rẻ. Menu đồ uống đa dạng. Nhân viên phục vụ thân thiện. Quán nằm trong hẻm rộng và có để bảng to ở đầu hẻm nên rất dễ tìm, chỗ để xe vô tư. Ở phía đối diện có quán rau câu dừa vs chè thái v.v có thể order sang ăn chung thoải mái. (*The pizza here is quite unique and different from the traditional style, so I do not like it much. The spaghetti with minced beef sauce and cream sauce was delicious. The food prices are affordable. The menu offers a variety of drinks. The staff is friendly. The restaurant is in a wide alley with a large sign at the entrance, making it easy to find. It is an available parking space. Across the street is a jelly and Thai dessert stall, so you can easily order from there and enjoy it together.*)

Question: Khách hàng đánh giá thế nào về món mỳ ý?(*How do customers review the spaghetti dish?*)

Answer: **Mỳ ý sốt bò băm và sốt kem rất ngon** (*The spaghetti with minced beef sauce and cream sauce was delicious*)

Sample 2

Review: Mua hôm 3/9 có một vài nhận xét như sau: màn hình sáng quá yếu phải bật độ sáng 80% trở lên, cam sau ok, cam trước sẽ không phù hợp với những người thích selfie, máy hơi dày và nặng, hình ảnh hiển thị **đẹp**.(*Bought on September 3rd, there are a few comments as follows: the screen brightness is too weak, it needs to be set at 80% brightness or higher, the rear camera is good, the front camera may not be suitable for selfie enthusiasts. The device is slightly thick and heavy, with beautiful display images.*)

Question: Chất lượng hình ảnh trên điện thoại như thế nào? (*How is the image quality on the phone?*)

Answer:đẹp (*beautiful*)

Table 2: Some examples of the machine reading comprehension task.

were held to reach a consensus.

- Every week, a meeting was conducted to discuss recorded cases, find solutions, and update the guidelines if necessary. Additionally, we randomly sampled 10% of the pairs created from each batch of cross-checking by the annotators to assess the quality of the cross-checking process. These samples were presented and discussed during the meeting. Finally, we randomly divided the ViRe4MRC into training, validation (val), and test sets, approximately in an 8:1:1 ratio.

3.2 Annotation Guidelines

This part presents the guidelines for creating question-answer pairs in ViRe4MRC. These guidelines are based on the UIT-ViQuAD (Nguyen et al., 2020a) corpus guidelines. Regarding the answer format, the answer must be a span in the review. The starting and ending characters of the answer should not be punctuation marks (such as periods, commas, or question marks) or spaces. In terms of content, the answer must accurately address the given question. Additionally, the answer should be the shortest possible since information extraction in MRC aims to find the exact answer to a question within a passage and avoids ambiguity. Moreover, having the shortest answer helps to eliminate redundant information and reduce the computational resources required for processing.

Inspired by Nguyen et al. (2020b), we divided question-answer into five reasoning types: word-matching (WM), paraphrasing (PP), single-sentence reasoning (SSR), multi-sentence reasoning (MSR) and ambiguous/insufficient (AoI). According to ViMMRC (Nguyen et al., 2020b) and UIT-ViQuAD (Nguyen et al., 2020a), PP, SSR, and MSR reasoning types have a higher proportion in the corpus than WM but lower results. Therefore, these types of reasoning are considered challenging. We encourage annotators to create more question-answer pairs of these types to maintain the level of challenge and difficulty for the corpus.

In the UIT-ViQuAD corpus (Nguyen et al., 2020a), the reasoning type of word matching only accounts for 13.35%, and in the ViMMRC corpus (Nguyen et al., 2020b), it accounts for

25.85%, but it has the highest accuracy among all reasoning types. Furthermore, in ViMMRC, Nguyen et al. (2020b) explained that the words in the question and the context match, models can easily provide an accurate answer. This demonstrates that word matching is an easy type of reasoning that can achieve high results without requiring a large amount of training data. Trischler et al. (2017) also suggests that WM reasoning is an easy type of reasoning. Therefore, we do not encourage annotators to create this reasoning with a high frequency.

3.3 Corpus Statistics

In this section, we present detailed information about the ViRe4MRC corpus. We analyze various aspects, such as the corpus size, the proportion of question types, types of reasoning, and the length of reviews and questions in the corpus. This analysis helps us configure experimental settings and support the analysis of results in the subsequent section.

Our corpus comprises 6,603 question-answer pairs extracted from 2,174 reviews. Specifically, there are 3,422 question-answer pairs in the smartphone domain and 3,181 in the restaurant domain. The difference in the number of question-answer pairs between the two domains is only 241 pairs (under 10%). This difference is not significant, which allows the models to learn without bias toward any specific domain. The result of each domain is presented in a Table 5.

	Entire	**Train**	**Val**	**Test**
Reviews	2,174	1,739	217	218
Question-answer pairs	6,603	5297	637	669
Max review length	205	205	184	173
Max question length	27	27	25	25
Max answer length	75	65	75	48

Table 3: Corpus statistics.

Table 3 presents an equivalent number of reviews and question-answer pairs in the validation and test sets, facilitating a fair evaluation of the models and ensuring accurate results. This equivalence enables an objective representation of the quality of the corpus. Furthermore, the max length of reviews and questions in the training, validation, and testing sets is 232, 209, and 198, respectively, as the basis for setting the maximum input length in our experiments.

Additionally, we assess the question types in the ViRe4MRC. Approximately 500 samples from the test set are randomly selected to estimate the proportion of question types. Subsequently, we manually classify the questions, including "how", "what", "how much/how many", "how long", "which", "why", and "other" allowing for the calculation of the quantity rate for each question type.

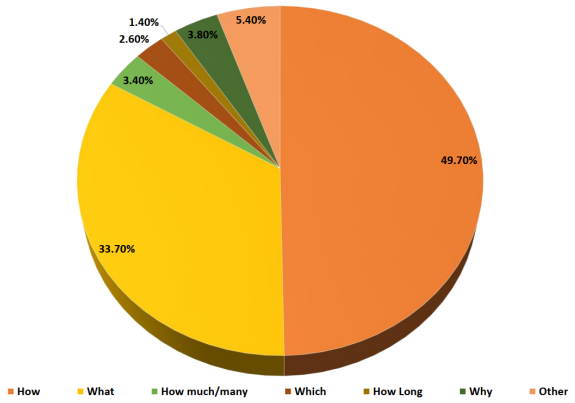

Figure 2: Proportion of question types.

Figure 2 illustrates that the **HOW** question type is the most frequent, aligning with our observations during data creation. We noticed that most reviews provided descriptive information about specific products or services, resulting in a higher occurrence of HOW questions.

In addition to calculating the proportion of question types, we evaluate the proportion of different types of reasoning. The categorization of the corpus is based on the proportions observed in 500 manually classified samples. As depicted in Figure 3, **paraphrasing (PP)** represents the **highest percentage**, followed by word matching (WM), multisentence reasoning (MSR), single-sentence reasoning (SSR), and ambiguous/insufficient (AoI).

4 Experiments and Results

4.1 Baseline Models

In this paper, we evaluate monolingual language models (ViBERT, PhoBERT, and vELECTRA) and multilingual language models (mBERT and XLM-R) on the ViRe4MRC benchmark.

- **BERT** (Devlin et al., 2019) is a powerful pre-trained model for many Nat-

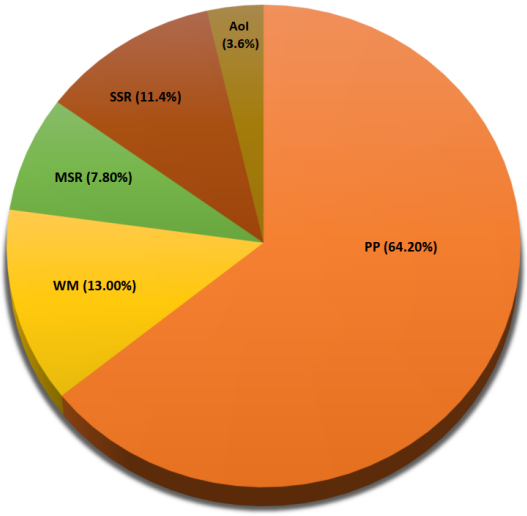

Figure 3: Proportion of reasoning types.

ural Language Processing (NLP) tasks. BERT comprises transformer encoder layers (Vaswani et al., 2017) (12 layers for BERT_{Base}, 24 layers for BERT_{Large}) and is pre-trained on a large corpus (the vocabulary size of mBERT is 30,000 tokens and number of tokens is 3.3B). In this study, we use mBERT.

- **XLM-R** (Conneau et al., 2020): This model is pre-trained on more than 100 languages (including Vietnamese) with a larger text corpus size compared to BERT's pre-trained multilingual corpus. For Vietnamese, XLM-R is trained on a large corpus (over 137GB and a number of tokens of over 24B).

- **ViBERT** (Bui et al., 2020): ViBERT is a BERT model but pre-trained in Vietnamese. The training data used for ViBERT is approximately 10GB, with a vocabulary size of over 38,000 tokens. The training process for ViBERT takes longer than training for vELECTRA.

- **PhoBERT** (Nguyen and Tuan Nguyen, 2020): PhoBERT is a monolingual pre-trained model for the Vietnamese language. The training data for PhoBERT comprises approximately 20GB, 3B tokens and a vocabulary size of 64,000 tokens. Similar to mBERT, PhoBERT comes in two versions: Base with 12 layers and Large with 24 layers.

Models	Validation		Test	
	F1(%)	EM(%)	F1(%)	EM(%)
XLM-R$_{Large}$	75.27	45.8	**78.13**	**44.25**
PhoBERT$_{Large}$	72.02	43.17	75.59	42.15
PhoBERT$_{Base}$	70.35	42.54	71.36	38.71
XLM-R$_{Base}$	66.76	38.15	67.37	36.92
mBERT$_{Base}$	58.28	30.46	63.15	33.48
ViBERT	55.73	28.73	54.30	26.76
vELECTRA$_{Base}$	50.93	26.22	54.21	27.06

Table 4: Model result on ViRe4MRC.

- **vELECTRA** (Bui et al., 2020): ELECTRA and vELECTRA have the same architecture; the difference is that ELECTRA is pre-trained in the Vietnamese language. vELECTRA is trained on 60GB of Vietnamese text. The vocabulary size of the corpus comprises over 32,000 tokens.

4.2 Evaluation Metrics

In this paper, we use two evaluation metrics: F1 and EM, as proposed by Rajpurkar et al. (2016) and used in Nguyen et al. (2020a)

4.3 Experimental Settings

In this study, we implemented several pre-trained models from Hugging Face[7], including mBERT, XLM-R, ViBERT, PhoBERT, and vELECTRA. For PhoBERT and XLM-R, each model has two versions: large and base. The large version has 24 transformer encoder layers, while the base version has 12. To distinguish between the versions, we added the "Large" or "Base" suffix to the model names. We made certain modifications to the hyperparameters.

Section 3.3 presented the maximum length of reviews and questions in the training, validation, and testing sets, which are 232, 209, and 198, respectively. Therefore, we used a maximum input length of 256. We conducted the models with the following settings: batch size = 32, 64; learning rate = 1.00E-04, 5.00E-05; epoch = 3, 5; Adam optimizer; Cross-Entropy loss function. We selected these hyperparameters based on our experience.

4.4 Results

Table 4 presents the results of review-based MRC models. The XLM-R_{Large} model achieves the best result with an F1 score of 78.13% and an EM score of 44.25%. The PhoBERT_{Large}

[7] https://huggingface.co/models

model followed with the F1 score of 75.59% and an EM score of 42.15%. The lowest-performing models were vELECTRA and ViBERT.

The XLM-R model achieves the best result. It undergoes training on large corpora encompassing multiple languages, including Vietnamese. It utilizes techniques to enhance results for languages with limited resources, such as Vietnamese. PhoBERT also delivers high results trained on Vietnamese data.

However, PhoBERT exhibits lower results than XLM-R, potentially due to the scale of the training data. Although both models are based on RoBERTa, XLM-R is trained on a larger and more diverse corpus, including 137GB of Vietnamese text, over 24 tokens compared to 20GB of text, 3B tokens of PhoBERT. This advantage enables XLM-R to handle user-created content more effectively.

5 Result Analysis

5.1 Result of Question Types

We randomly selected 500 samples from the test set and categorized them into different question types based on their distribution, as outlined in Section 3.3. Subsequently, we assessed the F1 and EM scores of the XLM-R$_{Large}$ models. Figure 4 presents results of question types.

The XLM-R$_{Large}$ model performed best for the "how much/many" question type. The "why" question type yielded the lowest accuracy, despite accounting for 3.8% (ranking fourth), "why" is higher than that of question types such as "how many/much", "how long", and "which". The "how" and "what" question types demonstrated significantly higher accuracy than the "why" questions. Thus, the "why" question is a difficult type.

5.2 Results of Reasoning Types.

We randomly select and classify 500 samples from the test set into different reasoning types based on their proportion, as outlined in Section 3.3. Subsequently, we assessed the F1 and EM scores of the XLM-RoBERTa$_{Large}$. Figure 5 shows several conclusions.

Figure 5 shows that the XLM-R$_{Large}$ model performs best on the WM type of reasoning. Difficult reasoning types such as PP, SSR, MSR, and AoI yield much lower results than WM. MSR has the lowest result with the EM metric, and AoI has the lowest F1 score. This experiment shows that reasoning types like PP, SSR, and MSR are challenging, while WM is an easier type of reasoning, as we presented in Section 3.2.

5.3 Result of Two Domains

We created the corpus in two domains of review domain: smartphones and restaurants. In this section, we compare the results of these two domains of the test set to determine the level of challenges in each domain. The XLM-R$_{Large}$ model achieved the highest results; hence, we utilized this model for the experiments. The results are presented in Table 5.

Models	**Smartphone**		**Restaurant**	
	F1(%)	EM(%)	F1(%)	EM(%)
XLM-R$_{Large}$	78.47	44.61	77.75	44.01

Table 5: The result of XLM-R$_{Large}$ with two domains.

Table 5 illustrates no significant difference in the results between the two domains. Therefore, both domains pose similar challenges due to their linguistic and stylistic similarities. Based on this experiment, it can be concluded that the principles outlined in Section 3.2 are rational, effective, and capable of generalizing across various customer review domains. Additionally, this experiment also demonstrates that there is no bias towards any specific domain, as presented in Section 3.3.

5.4 Error Analysis

In Table 4.4, the XLM-R$_{Lare}$ model demonstrated the highest result. As a result, we analyzed the errors encountered with this model to gain a deeper understanding of the ViRe4MRC corpus. We randomly selected 150 samples for each model from the test set where the models were mispredicted (EM=0). We classified the errors according to the classification system of Wadhwa et al. (2018). Figure 6 shows the proportion of different error types.

- **Soft correct error**: This error can be addressed by introducing more diverse and acceptable answers to the questions without relying on pre-existing answers. The UIT-ViQuAD (Nguyen et al., 2020a) utilizes this approach. Additionally, the

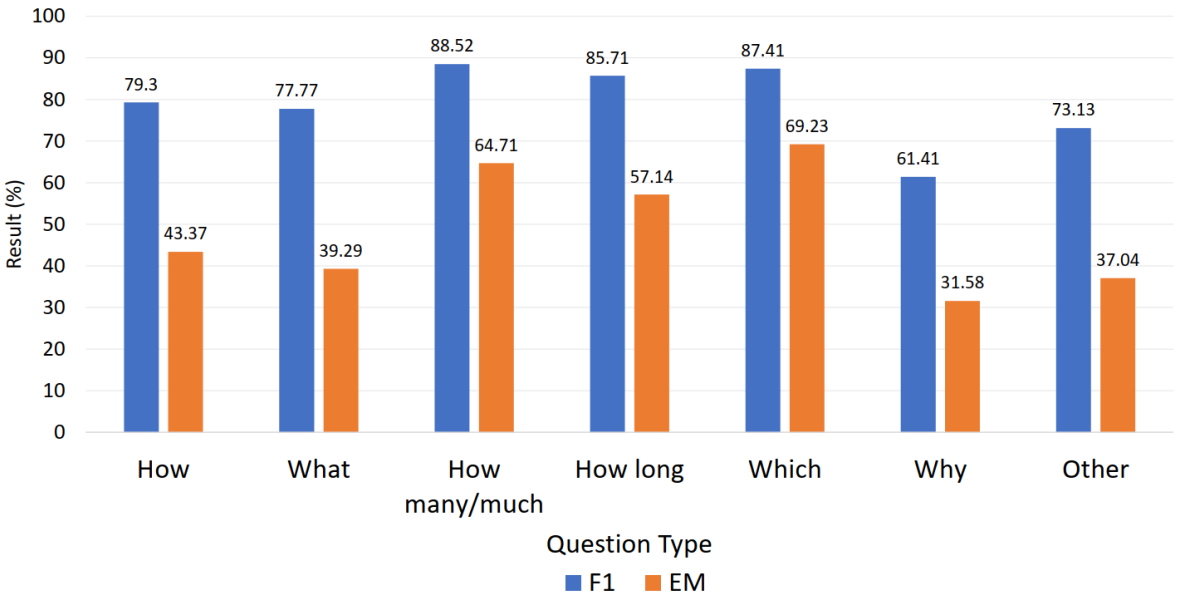

Figure 4: The result of XLM-R$_{Large}$ with question types.

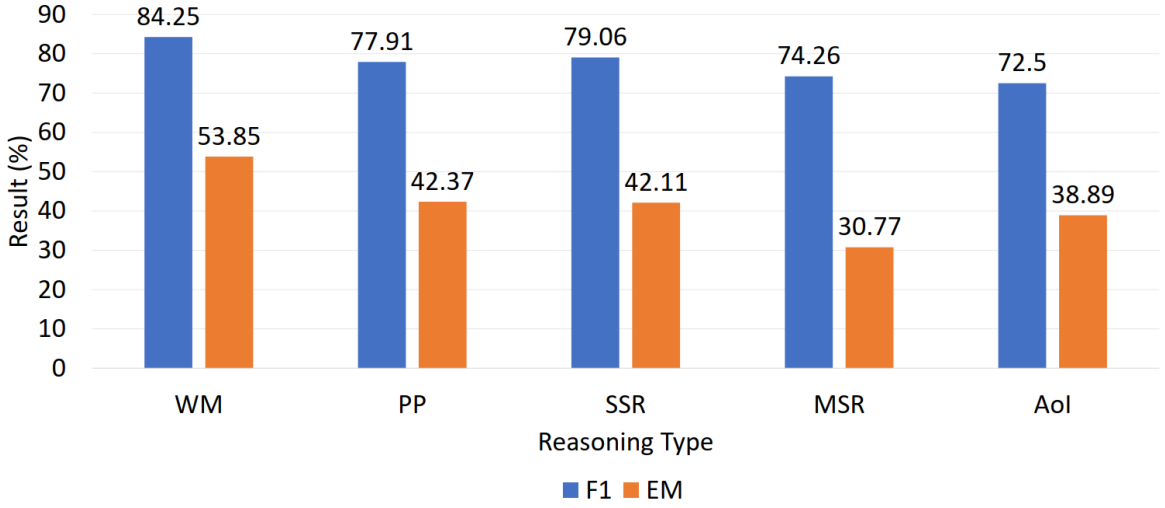

Figure 5: The result of XLM-R$_{Large}$ with the reasoning types.

soft correct error may be caused by errors in the data creation process or inadequate strictness in data creation rules. Reanalyzing the question-answer pairs provided by annotators may help identify and address the underlying issues.

- **Incorrect answer boundary (longer/shorter) error**: This error may arise when the model struggles to handle sentences containing acronyms, spelling mistakes, or breaks that do not follow formal language conventions. Such difficulties in understanding the language and determining the boundaries of ideas within sentences can lead to errors in predicting answer boundaries.

- **Paraphrase error**: This error category encompasses cases where the question paraphrases specific sections of the language being inquired about, making it challenging to match lexical patterns and resulting in inaccurate predictions.

- **Same entity type error**: In the reviews, multiple answers are scattered across paragraphs related to the same issue. Selecting the correct answer for a question becomes

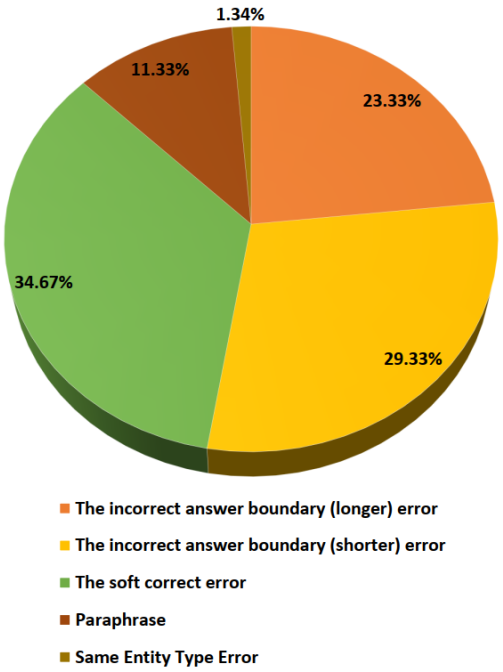

Figure 6: The proportion of reasoning type error.

problematic in such cases. This error occurs because annotators ask questions with multiple answers located far apart within the review during the data creation process.

6 Conclusion and Future Directions

In conclusion, we successfully created the first Vietnamese corpus (ViReMRC) for review-based machine reading comprehension. The corpus comprises 6,603 question-answer pairs from 2,174 reviews. Furthermore, we implemented deep learning models, including mBERT, PhoBERT, XLM-R, vELECTRA, and ViBERT, and evaluated them with our corpus. The XLM-R$_{Large}$ model achieved the best results, with 44.25% EM and 78.13% F1 on the test set. ViRe4MRC is an exciting and challenging corpus that contributes to diversifying the data sources for the MRC field and serves as a benchmark for future research in the customer reviews domain.

In future work, firstly, we plan to expand the corpus size by adding more data to the corpus to enable the machine to learn from a broader range of questions and answers and address any data errors that may arise. Secondly, we intend to collect and develop additional data from other domains to create an open-domain corpus that can enhance the ability of the model to handle diverse topics. Thirdly, we explore further data pre-processing techniques to effectively handle informal data types that often contain punctuation and spelling errors. Lastly, we aim to integrate a query step to create a retriever-reader question-answering (QA) system (Van Nguyen et al., 2023) and generative language models (Tran et al., 2021; Phan et al., 2022) suitable for real-world applications.

Acknowledgement

We thank San Qui Vu for supporting this paper. This research is funded by the University of Information Technology-Vietnam National University HoChiMinh City under grant number D1-2023-40.

References

Anonymous. 2021. Viqa-covid: Covid-19 machine reading comprehension dataset for vietnamese. In *ACL ARR 2021*.

The Viet Bui, Thi Oanh Tran, and Phuong Le-Hong. 2020. Improving sequence tagging for Vietnamese text using transformer-based neural models. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 13–20, Hanoi, Vietnam. Association for Computational Linguistics.

Danqi Chen. 2018. *Neural reading comprehension and beyond*. Stanford University.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Tin Van Huynh, Kiet Van Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Sentence extraction-based machine reading comprehension for vietnamese. In *Knowledge Science, Engineering and Management*, pages 511–523, Cham. Springer International Publishing.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *ArXiv*, abs/1704.05179.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 325–332, College Park, Maryland, USA. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Khang Le, Hien Nguyen, Tung Le Thanh, and Minh Nguyen. 2022. VIMQA: A Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6521–6529, Marseille, France. European Language Resources Association.

Wendy Grace Lehnert. 1977. *The Process of Question Answering.* Ph.D. thesis, USA. AAI7728146.

Son T. Luu, Mao Nguyen Bui, Loi Duc Nguyen, Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Conversational machine reading comprehension for vietnamese healthcare texts. In *Advances in Computational Collective Intelligence*, pages 546–558, Cham. Springer International Publishing.

Son T Luu, Khoi Trong Hoang, Tuong Quang Pham, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. A multiple choices reading comprehension corpus for vietnamese language education. *arXiv preprint arXiv:2303.18162*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020a. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kiet Van Nguyen, Khiem Vinh Tran, Son T. Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 136–142.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Triet Thai, Ngan Chu Thao-Ha, Anh Vo, and Son Luu. 2022. Uit-vicov19qa: A dataset for covid-19 community-based question answering on vietnamese language. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 801–810.

Kim Nguyen Thi Thanh, Sieu Huynh Khai, Phuc Pham Huynh, Luong Phan Luc, Duc-Vu Nguyen, and Kiet Nguyen Van. 2021. Span detection for aspect-based sentiment analysis in Vietnamese. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 318–328, Shanghai, China. Association for Computational Lingustics.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2021. Bartpho: pre-trained sequence-to-sequence models for vietnamese. *arXiv preprint arXiv:2109.09701*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Nguyen Van Kiet, Tran Quoc Son, Nguyen Thanh Luan, Huynh Van Tin, Luu Thanh 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).

Kiet Van Nguyen, Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2023. Multi-stage transfer learning with bertology-based language models for question answering system in vietnamese. *International Journal of Machine Learning and Cybernetics*, 14(5):1877–1902.

Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2022. New vietnamese corpus for machine reading comprehension of health news articles. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Soumya Wadhwa, Khyathi Chandu, and Eric Nyberg. 2018. Comparative analysis of neural QA models on SQuAD. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 89–97, Melbourne, Australia. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

A Error types

Table 6 presents several examples for error types in Section 5.4.

Error type	Review	Question	True answer	Predicted answer
The soft correct error	. . . đi 2 lần ở quán này , gọi 2 thức uống và ráng uống ké mấy đứa bạn nhưng chưa lần nào thấy ngon mà phải nói theo cảm nhận của mình là quá **dở**. . . (. . . *went to this restaurant twice, ordered 2 drinks, and tried to drink with friends but never felt good, but in my opinion, it was too bad*. . .)	Đánh giá của khách hàng về các đồ uống là gì?(*What are customer reviews about the drinks?*)	**dở** (*bad*)	**quá dở** (*too bad*)
The incorrect answer boundary longer error	. . . Đồ ăn thì siêu mắc , ko tương thích với chất lượng . Ăn như đấm vào mồm , **phục vụ chậm** . . . (. . . Food is super-expensive, incompatible with quality. Eating like punch in your mouth, slow service . . .)	Các bạn nhân viên ở đây có tác phong thế nào? (How are the staff here?)	**phục vụ chậm** (slow service)	**Ăn như đấm vào mồm , phục vụ chậm** (Eating like punch in your mouth, slow service)
The incorrect answer boundary shorter error	.... Nhất là chụp đêm, màu trắng lại xinh và hợp với con gái nữa, bạn gái mình cũng rất hài lòng. . . (. . . Taking photos very nice. Especially at night, the white color is pretty and suits girls too, my girlfriend is also very satisfied. . .)	Đánh giá của người dùng về chất lượng camera của sản phẩm là gì? (What is the user rating of the camera quality of the product?)	Chụp hình max đẹp. Nhất là chụp đêm (Taking photos very nice. Especially at night)	Chụp hình max đẹp(Taking photos very nice)
Paraphrase	. . . Chip snap665 quá ổn để chiến game + pin trâu. **Đồ phân giải nói là HD+ nhưng ko tệ**. . . (. . . The snap665 chip is too good to fight games + battery life. The resolution says it's HD+ but not bad. . .)	Chất lượng đồ họa của điện thoại ra sao? (How is the graphics quality of the phone?)	**Đồ phân giải nói là HD+ nhưng ko tệ** (The resolution says it's HD+ but not bad)	**Chip snap665 quá ổn** (The snap665 chip is too good)
Same Entity Type Error	. . . Gía cả rất sinh viên nhé . . **chỉ từ 17k trở lại** thôi. . . (. . . The price is very cheap. . only from 17k and back. . .)	Giá cả ở quán ra sao?(How are the prices at the restaurant?)	**chỉ từ 17k trở lại** (only from 17k and back)	**rất sinh viên** (very cheap)

Table 6: Several examples for error types.