VlogQA: Task, Dataset, and Baseline Models for Vietnamese Spoken-Based Machine Reading Comprehension

Thinh Phuoc Ngo[1,2], **Khoa Tran-Anh Dang**[1,2],
Son T. Luu[1,2], **Kiet Van Nguyen**[1,2], **Ngan Luu-Thuy Nguyen**[1,2]
[1]Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{19520981,19520629}@gm.uit.edu.vn, {sonlt, kietnv, ngannlt}@uit.edu.vn

Abstract

This paper presents the development process of a Vietnamese spoken language corpus for machine reading comprehension (MRC) tasks and provides insights into the challenges and opportunities associated with using real-world data for machine reading comprehension tasks. The existing MRC corpora in Vietnamese mainly focus on formal written documents such as Wikipedia articles, online newspapers, or textbooks. In contrast, the VlogQA consists of 10,076 question-answer pairs based on 1,230 transcript documents sourced from YouTube – an extensive source of user-uploaded content, covering the topics of food and travel. By capturing the spoken language of native Vietnamese speakers in natural settings, an obscure corner overlooked in Vietnamese research, the corpus provides a valuable resource for future research in reading comprehension tasks for the Vietnamese language. Regarding performance evaluation, our deep-learning models achieved the highest F1 score of 75.34% on the test set, indicating significant progress in machine reading comprehension for Vietnamese spoken language data. In terms of EM, the highest score we accomplished is 53.97%, which reflects the challenge in processing spoken-based content and highlights the need for further improvement.

1 Introduction

Machine reading comprehension (MRC) is a natural language processing (NLP) task that requires machines to comprehend a given context to answer a question (Baradaran et al., 2022). Although there are numerous datasets available for MRC tasks in English (Dzendzik et al., 2021), existing datasets for reading comprehension tasks in Vietnamese are relatively limited and they have primarily focused on written documents, such as Wikipedia articles, textbooks, and online news articles. Spoken language represents an important and distinct domain that has not been fully explored. Spoken language exhibits unique characteristics such as slang, regional variations, and informal grammar structures that can present significant challenges for machine learning models. As a result of that, reading comprehension tasks that involve spoken language, which is closer to everyday language, require a different type of dataset.

To address this need, we introduce VlogQA - a new Vietnamese spoken language corpus for reading comprehension tasks originating from transcripts of YouTube vlogs. As a global online video-sharing and social media platform, YouTube provides a vast amount of spoken language data in natural settings. It is now the second-most visited website[1] in the world (and Vietnam) and the second-biggest social network with over 2.5 billion monthly users[2]. Starting with YouTube, we aim to establish a solid foundation and gain insights into language patterns. This initial experiment serves as a stepping stone, and if successful, additional platforms catering to diverse audiences can be subsequently incorporated into further research. The dataset contains 10,076 manually annotated question-answer pairs based on 1,230 transcript documents extracted from YouTube videos. Besides, we provide several baseline models and evaluate them on our new dataset to test the ability of computers to understand the spoken text in Vietnamese. Overall, this paper makes the following contributions:

- We introduce VlogQA, a new Vietnamese corpus for MRC tasks that focuses on natural spoken language. The corpus contains transcripts from videos covering the topics of food and travel and has a noticeably larger average transcript length compared to the context size of other similar datasets. The inclusion of spoken language data enhances the value of the

[1] https://www.similarweb.com/top-websites/

[2] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

corpus, making it an invaluable resource for research purposes. Additionally, this resource has the potential for developing and evaluating spoken language QA systems that leverage speech-to-text tools to extract information from recordings or live-stream videos. For instance, the corpus can facilitate the training of a QA system tailored for meeting recordings, thereby simplifying content extraction by obviating the need for extensive note-taking or traditional Meeting Minutes.

- We provide the creation process with sufficient annotation steps to assure the quality of the corpus. Besides, we conduct the analysis and comparisons regarding the corpus, including the number of question-answer pairs, length-based statistics, and the distribution of question types to get insight into the natural spoken language in Vietnamese. We choose UIT-ViQuAD - a pilot MRC corpus constructed on Vietnamese Wikipedia Texts, to perform a comprehensive comparative analysis for exploiting the characteristics of spoken language.

- Finally, we evaluate the performance of multiple transformer-based language models on the corpus and analyze their performance for the MRC task on the spoken language domain. From the empirical results, we identify certain constraints within the dataset and highlight areas that can be improved in future studies.

The paper is structured as follows. Section 2 discusses existing studies. Section 3 is about corpus creation and its statistics. While Section 4 presents information about the language models to be used; the experimental results of human and language models, plus error analysis on the corpus are presented in Section 5. Finally, Section 6 provides conclusions and directions for future work.

2 Related Works

UIT-ViQuAD (Nguyen et al., 2020a) is a span-detection dataset for the Machine Reading Comprehension (MRC) task in Vietnamese, containing 23,074 questions on 5,109 passages acquired from Vietnamese Wikipedia articles. This dataset is widely used as a benchmark in Vietnamese MRC research and has facilitated innovations in the field. Its later version, UIT-ViQuAD 2.0 (Kiet et al., 2022), includes 9,217 additional unanswerable questions, which addresses a limitation of extractive MRC models that struggle to identify answers that are not explicitly mentioned in the text. Building upon the foundation of UIT-ViQuAD, UIT-ViWikiQA (Do et al., 2021) is a sentence-detection dataset converted from UIT-ViQuAD and is designed for tasks that focus on sentence-level comprehension. In the health domain, ViNewsQA (Van Nguyen et al., 2022) is a dataset comprising 22,057 questions on 4,416 online health articles from a popular newspaper in Vietnam.

Apart from span-detection datasets, there are other types of question-answering datasets available. ViMMRC (Nguyen et al., 2020b) is the first Vietnamese multiple-choice QA dataset, containing 2,783 four-choice questions based on 417 reading passages from Vietnamese literature textbooks. The second version of ViMMRC (Luu et al., 2023) introduces 699 reading passages and 5,273 questions with variable numbers of choices. UIT-ViCoV19QA (Thai et al., 2022) utilizes online FAQ documents from trusted healthcare organizations to address COVID-19-related questions, and is introduced as the first community-based QA dataset in Vietnamese with a total of 4500 questions. ViMQA (Le et al., 2022) is a Wikipedia-based multi-hop dataset that provides over 10,000 questions designed to challenge models to perform complex multi-hop reasoning tasks, requiring them to refer to multiple evidence passages and perform explainable reasoning.

The availability and diversity of quality question-answering datasets are essential for the development of effective machine-learning models for natural language processing tasks. Spoken SQuAD (Lee et al., 2018b) is an English dataset that targets spoken content comprehension in the context of Wikipedia articles. It is derived from SQuAD (Rajpurkar et al., 2016) and employs text-to-speech tools to generate the spoken context. Similarly, the ODSQA (Lee et al., 2018a) dataset focuses on spoken data and is based on the Delta Reading Comprehension Dataset (DRCD) (Shao et al., 2018), a Chinese contains 30.000+ questions from 2,108 Wikipedia articles. However, unlike Spoken SQuAD, ODSQA's audio is generated by humans.

In summary, current Vietnamese MRC datasets have mainly concentrated on formal types of content, such as Wikipedia articles, textbooks, and online news articles. While there are spoken-based

Question	Transcript	Answer
Nên chọn thịt như thế nào để không bị khô và vẫn giữ được độ mềm? *(What type of pork should be selected to avoid dryness while maintaining its softness?)*	[...] thật này mình sẽ xào cho nó chính nhé thật này nó có **vừa nạc vừa mõ** đó các bạn linh chi Mần ăn thì nó sẽ có cái độ mềm mềm béo nhá chứ mình làm không mấy thì ăn nó rất khô [...] *(we stir-fry the meat until it's really done the meat should be fatty meat type When being cooked it will have a tender texture otherwise it will be dry)*	"answer_start": 2196, "text": "vừa nạc vừa mỡ" *(fatty meat)*
Vì sao điểm khảo cổ Sa Huỳnh phải đổi tên? *(Why did the Sa Huynh archaeological site have to change its name?)*	[...]lần đầu được tìm thấy vào năm 1909 bởi nhà khảo cổ học người Pháp Venus trước đây Nghĩa danh này có tên là sao Hoàng tức là nhạc vàng song vì chữ hoàng lại **trùng tên Với Chúa Nguyễn Hoàng** cho nên đọc lái lại là thành sa huỳnh [...] *(first discovered in 1909 by a France arrchaeologist, Vinet. the site once had a name Sa Hoang which means golden sand however Hoang is the same name as Lord Nguyen Hoang, so it had to be euphemized to Sa Huynh)*	"answer_start": 893, "text": "trùng tên Với Chúa Nguyễn Hoàng" *((Hoang) is the same name as Lord Nguyen Hoang)*

Table 1: The examples in the corpus include ASR errors in Vietnamese, which are indicated by underlined text. The corresponding corrected English translations are also provided.

question-answering datasets available in other languages, such as Spoken SQuAD and ODSQA, they are still limited to Wikipedia content.

3 Corpus for Vlogs Reading Comprehension

3.1 Annotation Guidelines

Table 1 illustrates the structure of examples in the corpus, which is organized as a triplet *(q, t, a)*. We describe the reading-comprehension task in the scope of this paper as follows: Given a transcript document *t* of a Youtube vlog, one must comprehend and extract the answer *a* for the question *q*. The answer *a* must represent a specific word or phrase that is present in the transcript *t*.

Annotators play a vital role in ensuring the quality of the corpus by comprehending each transcript and creating at least five questions for it. If a transcript is too ambiguous or contains excessive ASR errors, annotators are advised to discard it. Similar to other single span-detection MRC datasets, the answer to a given question must be derived from the transcript's context and represent the shortest continuous meaningful phrase that matches the question. In addition, the answer must be a whole word or phrase. It is recommended that annotators generate the questions using their own words and include a diverse range of question types, answers, and supporting evidence.

3.2 Data Creation Process

The proposed process for creating the VlogQA corpus includes four main stages: Transcript collection, QA pair creation, Corpus modification, and Quality assurance. Figure 1 illustrates the overview of the creation process for the corpus and the detailed description is provided as follows.

3.2.1 Transcript collection

The transcripts in the corpus were collected from Vietnamese YouTube vlogs with topics related to food and cooking tutorials, travel, or both. The channels that own the vlogs should have a large subscriber base; in this dataset, we set the minimum number of subscribers at 200,000 to ensure that the content is acceptable and relevant to a portion of the community. For each vlog, the transcript was collected using a Python API[3] that returns a list of short speech-span transcriptions and is later combined into a single document. In this paper, the transcripts were kept in their original size and not segmented into smaller passages.

3.2.2 QA pair creation

Corresponding to each transcript document, one annotator is asked to read, comprehend and then create question-answer pairs following the annotation guidelines. Having completed this stage, the questions are collected and randomly chosen to form a set of 100 questions. This set is used to estimate the degree of agreement among annotators.

3.2.3 Corpus modification

To improve the consistency of the annotator and ensure corpus validity, the annotators are tasked

[3] https://pypi.org/project/youtube-transcript-api/

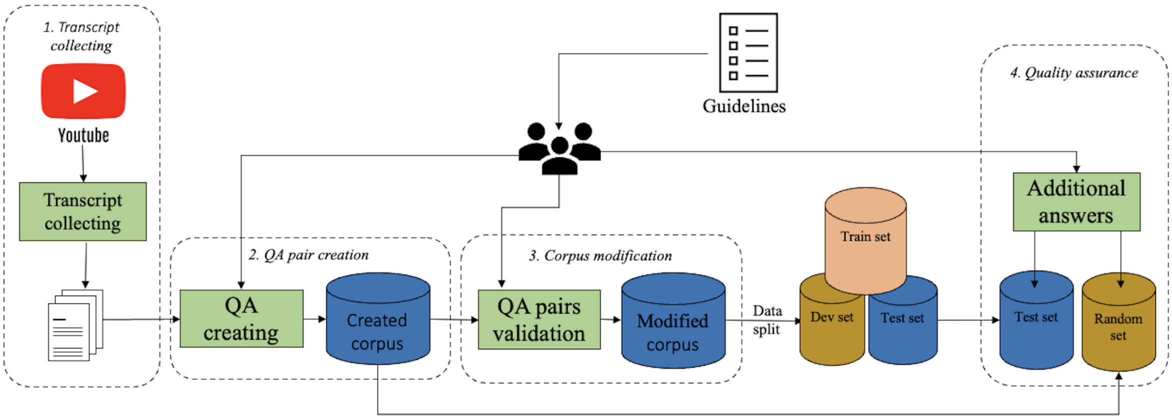


Figure 1: The overview process of creating the corpus.

with conducting the following 3 steps: (1) Additional training: annotators participate in additional training to better understand the evaluation criteria and guidelines. (2) Self-validation: annotators do a self-check of their own work to identify and correct any errors or inconsistencies, such as unclear questions, incorrect answers, lack of information questions, and incorrect boundary answers. (3) Cross-annotator-validation: the self-checked data is later reviewed by another annotator to ensure accuracy and consistency. The modified dataset is divided into three subsets: train, test, and development with a ratio of 8:1:1 based on the question. Each transcript is assigned to only one subset.

3.2.4 Quality assurance

To determine the reliability of the corpus, we perform the following two examinations:

1. **Inter-rater agreement:** This step aims to estimate the quality of the annotators' work. Each annotator independently provides an additional answer for each question in the random set. During the process, annotators work without referring to the corpus's answers. To estimate the inter-rater agreement, a measure of the degree to which annotators agree on their labels, we employ three metrics: Cohen's Kappa (Cohen, 1960), Fleiss' Kappa (Fleiss, 1971), and Krippendorff's alpha (Krippendorff, 2013). Additionally, we also calculate the overlap among answers using ROUGE metrics (Lin, 2004), and compute the semantic similarity of answers among annotators by using BERTScore (Zhang et al., 2019).

2. **Human performance:** After splitting the corpus, an independent team is enrolled to augment the test set with additional answers. The F1 score and exact-match metrics are used to evaluate human performance on the dataset.

3.3 Dataset Analysis

3.3.1 Overall statistics

Inferred from Table 2, the dataset comprises 1,230 vlog transcripts, of which only 64 transcripts are manually created by video creators; the remaining are generated automatically by Youtube. As shown in Table 3, most of the transcripts have less than 5,000 words, the shortest transcript consists of 223 words while the longest one has 38,228 words.

Vietnamese relies heavily on word order and function words to convey meaning and express grammatical relationships, rather than inflectional affixes. Words in Vietnamese are constructed from syllables ("tiếng"), which are the basic unit of meaning, and words can be mono-syllabic or polysyllabic. Vietnamese is also known for its extensive use of compound words, which combine two or more words to create a new word with a distinct meaning (Binh, 2021). Segmentation is essential for identifying the tones of syllables in a word, which can affect the meaning of the word and the overall meaning of a sentence. However, the Vietnamese language lacks a standard for word segmentation (Nguyen et al., 2012). We use a Python Vietnamese toolkit[4] to segment words, following the methodology of the UIT-ViQuAD paper. We also re-calculate some statistics of UIT-ViQuAD v1.0, using the latest version of the tool to compare the two datasets.

[4] https://pypi.org/project/pyvi/0.1.1/

	VlogQA				UIT-ViQuAD			
	Train	**Dev**	**Test**	**Total**	**Train**	**Dev**	**Test**	**Total**
Context count	945	130	155	1,230	137	18	18	174
Question count	8,047	1,017	1,012	10,076	18,579	2,285	2,210	23,074
Avg. context length	2,789.5	2,779.5	2,498.9	2,751.7	153.7	148.8	155.8	153.4
Avg. question length	10.09	10.10	10.00	10.08	11.23	11.96	12.29	11.40
Avg. answer length	3.22	3.27	3.31	3.24	8.06	8.45	8.93	8.18
Vocabulary size	34,288	12,639	13,336	39,211	36,940	9,746	10,263	42,545

Table 2: Overall statistics of our dataset and UIT-ViQuAD.

In comparison to UIT-ViQuAD, which consists of 23,074 questions, our dataset is smaller, with 10,076 questions. On average, the length of the questions between the 2 datasets is not much different; however, our answers are significantly shorter, only 3.24, compared with 8.18 words per answer of UIT-ViQuAD. Our dataset used more context documents, a total of 1,230 transcripts compared with 174 passages. Additionally, the transcripts in our dataset are much longer on average, with an average length of 2,751.7 words, compared to the majority of UIT-ViQuAD's context passages ranging from 101 to 200 words.

Despite the difference in the number of questions, our dataset offers a vocabulary size of 39,211, which is only 7.83% less than UIT-ViQuAD with a vocabulary size of 42,545. In this study, the vocabulary is estimated based on the segmented words of the context documents. Of the two corpora, there are 13,647 overlapping words, and our corpus has a unique vocabulary of 25,564 words. The most frequent words and phrases in our dataset are related to unit measurements, linking words, padding words, and pronouns. Those are commonly used in everyday scenarios and may be considered informal or unlikely to appear in formal writing or contexts. In Appendix A.2, we provide further details on the differences in vocabulary between the two datasets and the methods we used to identify them using word clouds.

Length	**Count**	**Percentage**
0 - 2,000	4,446	44.1
2,000 - 4,000	3,315	32.9
4,000 - 6,000	1,674	16.6
6,000 - 8,000	492	4.9
9,000 - 39,000	149	1.5

Table 3: Transcript length distribution.

3.3.2 Duration-based analysis

The following information on video length is calculated based on a total of 1,221 videos, as not all videos were available at the time of statistics. The results in Table 4a reveal that the average length of the selected videos is 1,272.95 seconds (21.2158 minutes). The shortest video lasts 60 seconds (1 minute), while the longest video has a duration of 19,190 seconds (5.331 hours). On average, travel-related videos have longer durations than food-related videos. Table 4b further supports the finding that the majority of videos from food channels have a duration between 400 to 1,200 seconds, while the majority of traveling channel videos typically range from 1,000 to 3,000 seconds.

	Food	**Travel**	**Total**
Video count	565	665	1,230
Avg. length	721.86	1,914.91	1,272.95
Max. length	3,040	19,190	19,190
Min. length	173	60	60

(a) Video length statistics by category (in seconds).

	Length	**Count**	**Percentage**
Food	<400	61	9.28
	400 - 800	383	58.30
	800 - 1,200	160	24.35
	>1,600	53	8.07
Travel	<1,000	66	11.70
	1,000 - 2,000	291	51.60
	2,000 - 3,000	151	26.77
	>3,000	56	9.93

(b) Distribution of the video length (in seconds).

Table 4: Statistics of video duration.

3.3.3 Inter-rater agreement

After the first annotation round (Section 3.2.2), we calculate the inter-rater agreement among six annotators on three metrics. However, given that Cohen's Kappa (Cohen, 1960) is designed for two annotators, we calculated the average degree of Cohen's Kappa agreement among all possible pairs of annotators. The average level of inter-rater agreement, as demonstrated by the results in Table 5, is approximately 0.44. This level of agreement falls within the moderate range (Landis and Koch, 1977).

The results of the ROUGE metrics (Lin, 2004) in Table 5 are significantly higher than the agreement degrees, suggesting that mismatches were mainly due to non-essential terms rather than fundamental disagreement on the answer. Although the annotators had captured the context, it also highlights that the Corpus modification stage should focus on improving the consistency of the annotators to ensure the reliability and validity of the corpus. Besides, the BERTScore (Zhang et al., 2019) value shows that the answers among annotators are significantly similar, ensuring high agreement between annotators.

Metric	Score
Cohen's Kappa (average)	0.4393
Fleiss' Kappa	0.4387
Krippendoff's Alpha	0.4398
RougeL	0.7672
Rouge1	0.7683
Rouge2	0.6776
BERTScore	0.8867

Table 5: Inter-rater agreement degree.

3.3.4 Question type analysis

We categorize the questions into seven types, namely Who, What, When, Where, Why, How, and Others. Additionally, the How-type questions are further divided into two subtypes: quantity-related questions, which inquire about the amount or number of something, and quality/method-related questions, which focus on the characteristics or techniques involved. The question labeling process is done manually because the diversity of question words in Vietnamese makes it hard to automate the process. For example, the English question word "when" can be translated into various Vietnamese question words, such as "khi nào", "lúc nào", "bao giờ", and others, depending on the context. The word "nào" occurs in many of these translations, but applying rule-based methods is difficult because "nào" can also mean "what/which" in other contexts. According to the statistics presented in Table 6, the distribution of question types in our dataset is different from that of UIT-ViQuAD. Although the proportions of the "What" type questions are similar in both datasets at 47.82% and 49.97%, our dataset has a larger proportion of questions of "How" type at 32.57%, compared to 9.09% in UIT-ViQuAD. This distribution of question types reflects the characteristics of the data domain, that food and travel content deliver large information about the quantity and it is easier for annotators to create questions of that type.

	UIT-ViQuAD (%)	VlogQA (%)
What	49.97	47.92
How	9.09	32.57*
Why	7.54	8.63
Where	5.64	5.25
When	8.96	3.35
Who	9.41	2.22
Others	9.41	0.07

Table 6: The proportions of question types in UIT-ViQuAD and VlogQA dataset. In the VlogQA dataset, the How-type is the sum of the How-quantity type (25.59%) and the How-quality type (6.98%), respectively.

4 Models for Reading Comprehension

Transformer (Vaswani et al., 2017) is a type of neural network architecture designed to process sequential data. In this paper, we carry out the MRC task and evaluate performance on the following group of transformer-based pre-trained language models:

- Multilingual language models, including (1) mBERT (Devlin et al., 2019) – an extension of BERT developed by Google, having been trained on over 100 languages, and (2) XLM-R (Conneau et al., 2020) – a Cross-lingual Model introduced by Facebook Research.
- Monolingual language models, including (3) PhoBERT (Nguyen and Tuan Nguyen, 2020), (4) BARTPho (Tran et al., 2022), (5) ViT5 (Phan et al., 2022) which are constructed on Vietnamese data.

The information about the size of pre-trained language models, the hyperparameter settings, and the environment for experiments are shown in Appendix A.3.

5 Empirical Results

5.1 Experimental results

In this section, we first present the experimental results of the language models and compare their performance with that of humans. The models are fine-tuned using the training and development sets.

Model	Dev (%)		Test (%)	
	EM	F1	EM	F1
mBERT	40.36	61.60	45.17	64.89
XLM-R$_{Base}$	45.78	65.63	47.73	68.71
XLM-R$_{Large}$	**51.41**	**72.39**	**53.97**	**75.34**
PhoBERT	23.32	35.49	23.06	34.37
BARTPho	18.91	30.27	20.43	32.58
ViT5	30.33	45.82	29.37	45.53
Human Performance	-	-	48.49	76.25

Table 7: Pre-trained language models performance on VlogQA test set in terms of EM and F1-score. The models were trained on the VlogQA corpus.

The results in Table 7 indicate that the XLM-R$_{Large}$ model outperforms the other models, achieving the highest scores in both EM (53.97%) and F1-score (75.34%). In contrast, PhoBERT's performance was even lower than that of the English pre-trained models, possibly due to its word tokenizing technique, which may not be optimal for handling the challenges posed by spoken language. Spoken language often contains various errors, stutters, and other linguistic features unique to spoken communication, making it a challenging task for natural language processing models like PhoBERT. The ViT5 is the best performance monolingual pre-trained on the VlogQA dataset, which is 29.37% for EM and 45.53% for F1-score. However, the results of ViT5 is not as good as XLM-R on the VlogQA dataset.

In order to assess human performance on the task, we computed the scores of two independent annotators on the test set. The resulting evaluation shows that the human performance achieved an EM score of 48.49% and an F1-score of 76.25%. Interestingly, the XLM-R$_{Large}$ model performed even better than humans on the EM metric, which is a remarkable accomplishment. However, on the F1-score, there is only a slight difference between the model and human performance. These findings suggest that the XLM-R$_{Large}$ model has the potential for this task, but there is still room for improvement in terms of F1-score.

Model	EM (%)	F1-score (%)
mBERT	16.67	37.05
XLM-R$_{Base}$	24.62	49.07
XLM-R$_{Large}$	**35.42**	**62.43**
PhoBERT	24.15	50.06
ViT5	8.28	19.30
BARTPho	2.07	12.21

Table 8: Pre-trained language models performance on VlogQA test set in terms of EM and F1-score. The models were trained on the UIT-ViQuAD corpus.

In addition to training transformer-based models on our training and dev sets, we also evaluate the performance exclusively on the UIT-ViQuAD training and dev sets and then test them on our test set. This is to evaluate whether the current pre-trained models are good when they were fine-tuned on another domain. The results of our evaluation, as shown in Table 7 and 8, indicate that XLM-R$_{Large}$ performs the best, but its performance decreases drastically when trained on the UIT-ViQuAD dataset. Surprisingly, the Vietnamese pre-trained model, PhoBERT, performs better when we train them on the UIT-ViQuAD dataset. However, it is still lower than the performance of the XLM-R. In general, the performance of the language model that was fine-tuned on the UIT-ViQuAD does not achieve the expected results for the MRC task on spoken text as it was fine-tuned on our VlogQA.

5.2 Error analysis

We exclude the "Others" question type in this section due to its negligible representation. Illustrated in Table 9 are the numbers of incorrect answers of each type and their proportions in the development set. An answer provided by the language model is considered wrong if the answer and the reference answer are not an exact match ($EM = 0$). Overall, the XLM-R$_{Large}$ model achieves superior performance compared to other models in all question types. Therefore, we will focus on analyzing the errors of the XLM-R$_{Large}$ model.

Based on the information in Table 9, the XLM-R model has the lowest error rate on Where and How (quantity) question types, at 33.96% and 34.93%. The What type questions make up the largest pro-

Question type	mBERT	XLM-R$_{Large}$	PhoBERT	BARTPho	ViT5
What	308 0.6123	248 0.4930	387 0.7694	408 0.8111	353 0.7018
How (quantity)	97 0.4236	80 0.3493	145 0.6332	159 0.6943	130 0.5677
How (method)	56 0.7887	49 0.6901	65 0.9155	65 0.9155	58 0.8169
Where	33 0.6226	18 0.3396	39 0.7358	38 0.7170	35 0.6604
Who	18 0.6923	12 0.4615	23 0.8846	23 0.8846	18 0.6923
Why	70 0.8434	62 0.7470	76 0.9157	78 0.9398	74 0.8916
When	18 0.5455	16 0.4848	28 0.8485	32 0.9697	25 0.7576

Table 9: The number and the rate of incorrect answers on the VlogQA development set, grouped by type, using the EM metric.

portion of our dataset, giving an error rate of 49.30%. The type that has the most error rate is the Why question type, with a rate of 74.70%.

The average F1 score of error predictions is 43.18%, and 73.46% of predictions have a non-zero F1 score. Common errors can be categorized as inconsistent identification of non-essential terms, which may result from the variability of the spoken language. Other common errors include misinterpretation of the nuances of the question, and providing completely wrong answers that are not supported by the information provided. We further provide examples of the errors in Appendix A.4.

6 Conclusion and Future Works

This paper presents VlogQA - a new Vietnamese reading comprehension corpus for spoken context. The corpus consists of 10,076 question-answer pairs generated by humans, sourced from 1,230 transcripts of YouTube vlogs. Each transcript has an average length of 2,752 words. In terms of question types, the dataset is predominantly composed of What-questions, accounting for 47.52% of the corpus. This is followed by How-questions, which make up 32.57% of the dataset. Other question types represented in the corpus include When, Who, and Why, among others. Our experimental results indicate that the annotation of the dataset is acceptably consistent, with an average inter-rater agreement of nearly 44%. The performance of the state-of-the-art multilingual model is comparable with humans in both F1-score and EM metrics; however, we believe that there is still room for improvement. In future work, we plan to enhance the corpus both in quality and quantity. We will explore techniques for improving the consistency of annotations and seek to expand the dataset with additional transcripts, spanning more topics. We also plan to augment the corpus with unanswerable questions, which will enable further exploration of machine capabilities.

Overall, this new Vietnamese reading comprehension corpus for spoken context provides a valuable resource for researchers and practitioners in the field of natural language processing. Moreover, We anticipate this dataset will facilitate advancements in Vietnamese language understanding and provide a benchmark for the evaluation of intelligent question-answering systems on human-spoken language. Furthermore, this corpus will enable the development of smart systems capable of retrieving valuable information from spoken language, thus contributing to the advancement of the field.

Limitations

Using spoken content as a data source ensures that the corpus reflects the diverse nature of spoken language and culture of everyday life, including informal settings. On the other hand, these distinct resources of Youtube also pose unique challenges for existing systems, including:

- **Accent and dialect:** Vietnamese is a tonal language with three main dialect regions (Northern, Central, and Southern), which means that there are differences in the way words are pronounced and used. This complexity and variation in real-life situations cause errors in automatic speech recognition (ASR) systems.

- **Audio quality:** Low-quality audio is difficult to transcribe accurately, leading to errors and inconsistencies in the dataset. Background noise, such as music or ambient sounds may interfere with the transcript quality, especially in outdoor recordings like travel vlogs.

- **Transcript format:** Unlike regular documents, the ASR system does not provide punctuation (e.g., commas and periods) or consistent letter cases (e.g., uppercase first letters in named entities), which may pose challenges for understanding the meaning of the transcript. Moreover, ASR transcripts do not support identifying speakers where multi-speakers are present.

- **Transcript length:** The length of vlogs in our dataset is highly variable, with some videos lasting under 10 minutes and others exceeding an hour, leading to the fact that most of the transcripts are significantly larger than the context provided by other datasets. There is a substantial amount of non-relevant information that needs to be filtered out to identify relevant information for each question.

These factors may put a negative impact on MRC systems. However, they also present opportunities to provide a unique dataset with vocabulary and word combinations specific to spoken language, which is rare in the existing datasets. Finally, we could not make our own pre-trained language model on spoken language text due to the limitation of computing resources such as GPU and memory. We hope the future pre-trained language models and large language models (LLMs) for spoken texts will improve the performance of the machine reading comprehension model for spoken language.

Ethics Statement

We select videos that are published and verified by YouTube, 94.80% of the transcript documents are automatically generated by YouTube's speech recognition. We keep all selected transcripts in their original form, and they are available at the time of collection. For the data annotation process, all annotators are supported with adequate remuneration for their work. The information about annotators is made anonymous.

Acknowledgments

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

References

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Ngo Binh. 2021. *Vietnamese: An essential grammar*. Routledge, Taylor & Francis Group.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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.

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.

Daria Dzendzik, Jennifer Foster, and Carl Vogel. 2021. English machine reading comprehension datasets: A survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Nguyen Kiet, Tran Son, Nguyen Luan, Huynh Tin, Luu Son, and Nguyen Ngan. 2022. Vlsp 2021-vimrc challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(2).

Klaus Krippendorff. 2013. *Content Analysis*. SAGE.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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.

Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-yi Lee. 2018a. Odsqa: Open-domain spoken question answering dataset. pages 949–956.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018b. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. pages 3459–3463.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

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.

Quy T. Nguyen, Ngan L.T. Nguyen, and Yusuke Miyao. 2012. Comparing different criteria for Vietnamese word segmentation. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 53–68, Mumbai, India. The COLING 2012 Organizing Committee.

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, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

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.

Chih Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset.

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, Manila, Philippines. De La Salle University.

Nguyen Luong Tran, Duong Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proc. Interspeech 2022*, pages 1751–1755.

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).

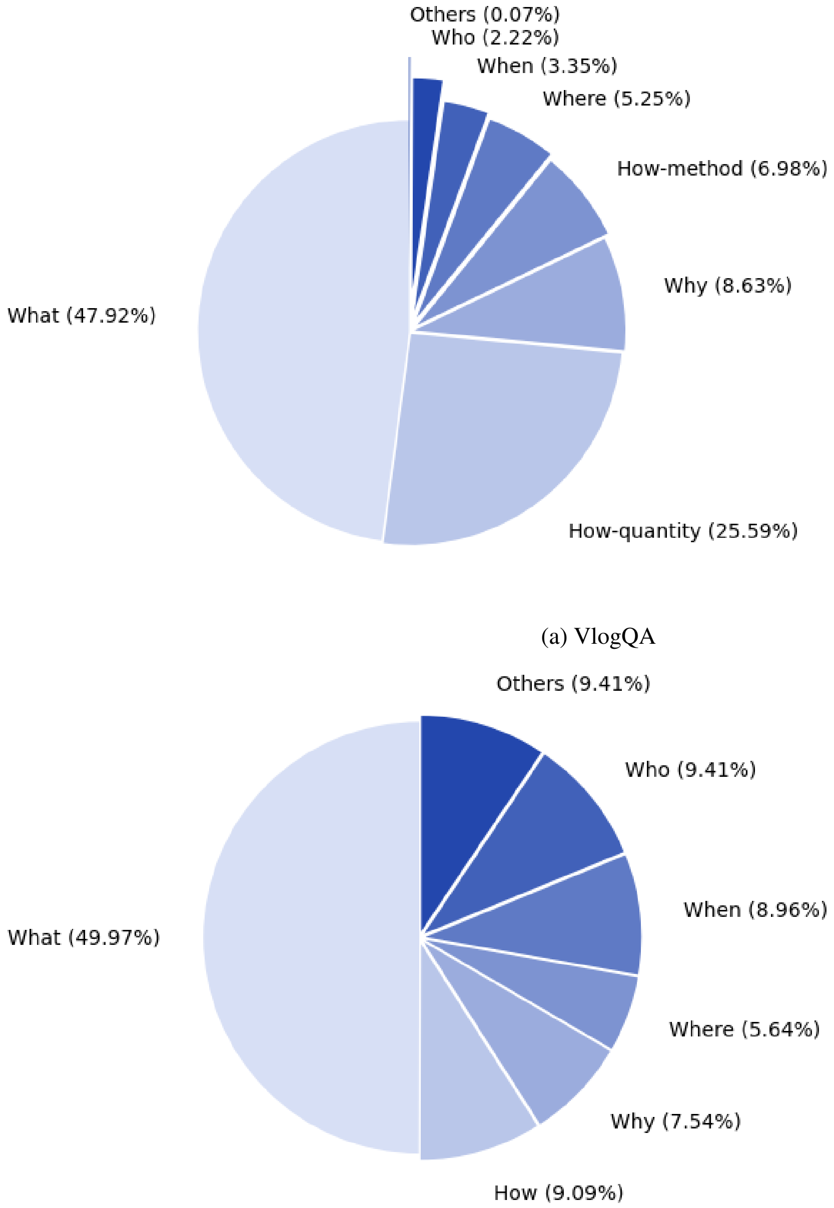
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendices

A.1 Question type distribution

Figure 2 presents pie charts showing the distribution of question types in the VlogQA and UIT-ViQuAD, in addition to Section 3.3.4.



(a) VlogQA

(b) UIT-ViQuAD

Figure 2: Question types proportions of the VlogQA and UIT-ViQuAD

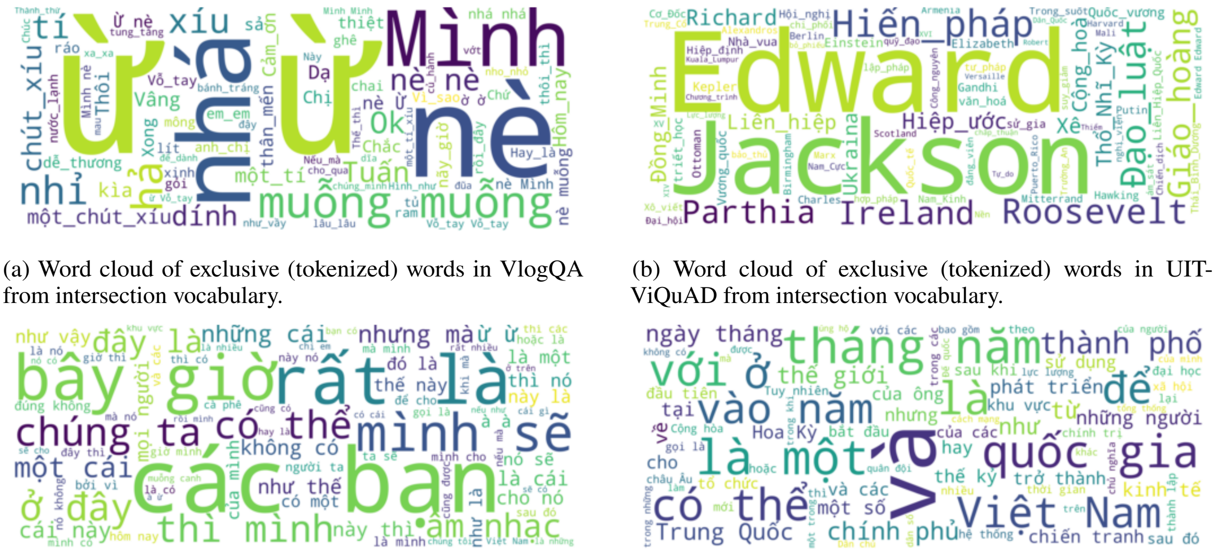
A.2 Vocabulary

Figure 3 shows the word clouds for the context documents in our dataset and UIT-ViQuAD. Each cloud is limited to 80 words, and we have opted not to apply any stop-word filters in these visualizations to preserve the essence of spoken materials.

As discussed in Section 3.3.1, our analysis revealed that the two corpora share an estimated 13,647 words. To further explore the distinctive vocabulary of each corpus, we created Figure 3a to display a visualization of the exclusive vocabulary in our corpus, which does not overlap with the shared vocabulary.

The most frequent words in our exclusive cloud pertain to padding words in real-life spoken language that are eliminated in written contexts, such as "ừ", "nhé", and "nè". It also includes the pronoun "mình", which is common in informal contexts and is similar to "I", "we", "me", and "us" in English. Similarly, Figure 3b presents a word cloud that showcases the unique words in UIT-ViQuAD. To generate these clouds, we tokenized the context documents to involve the estimated vocabulary.

The remaining part of Figure 3 presents the word clouds of the entire context documents of the two corpora. The UIT-ViQuAD cloud represents more formal words that frequently occur in informative contexts, such as "chính phủ", "tháng năm", "quốc gia", and "Việt Nam". On the other hand, VlogQA introduces a set of spoken language words, such as "rất là" to express emphasis on an adjective or adverb, or "các bạn", which is somewhat equivalent to "you guys" in English.



(a) Word cloud of exclusive (tokenized) words in VlogQA from intersection vocabulary.

(b) Word cloud of exclusive (tokenized) words in UIT-ViQuAD from intersection vocabulary.

(c) Word cloud of VlogQA (without tokenization).

(d) Word cloud of UIT-ViQuAD (without tokenization).

Figure 3: Word cloud presentations of VlogQA and UIT-ViQuAD.

A.3 Experimental settings

We train our baseline models in a Google Colab environment with a single NVIDIA Tesla T4 GPU. The pre-trained models are fine-tuned using our dataset under the default settings of HuggingFace Trainer API[5], *batch_size* = 8 and *epochs* = 10. We also set the *max_length* of the tokenizer to 512 (except for the case of PhoBERT, which is 256 due to the hardware limitations). The number of parameters for each model is described in Table 10. The baseline code is available at https://github.com/sonlam1102/vlogqa.

Model	**#parameters**
mBERT	179M
XLM-R (base)	279M
XLM-R (large)	561M
PhoBERT	135M
BARTPho	132M
ViT5	310M

Table 10: Number of parameters for empirical models

[5] https://huggingface.co/docs/transformers/

A.4 Incorrect prediction examples

Table 11 illustrates a sample of incorrect predictions for each question type, derived from the XLM-R$_{Large}$ model on the VlogQA development set. The observed discrepancy between the model's F1 score and EM performance may be attributed to the extraneous or insufficient usage of non-essential terms, which is why we selected these particular examples.

For the What-type question, the model's prediction may include an optional excessive term, "region", but eliminating it does not significantly impact the results in the Vietnamese aspect.

In the Quantity-type question, the term "dưới" (which means "under" in English) must be included in the answer to determine the constraint on the length of hair of Phu Quoc dogs. While the model's answer may be contextually relevant, it is not entirely accurate without the inclusion of the term "dưới".

In the case of the How-quality question, the model eliminated a term, which resulted in a minor grammatical flaw in Vietnamese (our translation to English may be deficient to fully reflect this example as the difference in the position of words in the 2 languages). It is worth noting that the model is extractive for spoken-based language, and in some contexts within this dataset, such a prediction/answer may still be acceptable despite the small error. However, our translation to English may also be deficient to fully reflect this example as the difference in the position of words in the two languages could affect the accuracy of the translation.

In the Why-type question, the model's prediction has correctly identified the context but may be missing some minor terms, as seen in previous examples. This elimination makes it more likely to be an answer to a What-type question in Vietnamese.

For the Where-type question, the model's prediction includes redundant terms that are not relevant to the question, which only seeks information about the location, not the designated name of it, the land of martial arts. While the model's answer may be partly correct, the inclusion of these unnecessary terms could potentially confuse the reader or listener and detract from the accuracy of the answer.

In the Who-type question, the model should not include terms that express an extreme as the question does not focus on this. The model's prediction should only include information that is relevant to answering the question and avoid adding unnecessary or extraneous information.

The illustrative examples provided do reflect the difficulties in processing Vietnamese spoken-based materials, particularly due to the complex grammar system and variations in pronunciation, intonation, and word order. While the questions in this dataset may not be considered hard, they can still be challenging for natural language processing models to accurately interpret and respond to. It is important to carefully consider the limitations of these models and the context in which they are being used when analyzing their performance on language tasks in Vietnamese.

	Question	Transcript	Reference	Prediction
What	Khu vực Bắc Ninh từng được được gọi bằng tên gọi gì?	những cái khu chợ xưa ấy thì xung quanh đây là những cái cột đá cổ Nhưng cái xà thanh xà bằng gỗ và mà ấy máy là mái ngói mái ngói theo đúng kiểu đặc trưng của vùng Bắc Bộ ngày xưa khu Bắc Ninh là gọi là **khu Kinh Bắc** nhé	Kinh Bắc	khu Kinh Bắc
	What was Bac Ninh called in the past?	*The old marketplaces had these ancient stone pillars around them. The wooden rafters and tiled roofs were of the typical style of the (classic Vietnamese) North. In the past, the region (around) Bac Ninh was called* ***Kinh Bac region***	*Kinh Bac*	*Kinh Bac region*
How (quantity)	Lông chó Phú Quốc dài bao nhiêu?	thứ nhất là nó phải lông **dưới 2cm** cái Lông nó sát gọi nó là quan sát cái thứ 2 là nó cái anh em của nó và nó gọi cái giới khoa học Tao gọi là nó có cách mạng bàn chân phát triển như là chân vịt	dưới 2cm	2cm
	What is the length of Phu Quoc dog's hair (in general)?	*Firstly, their hair must be* ***under 2cm*** *which means it is short Secondly, the researchers observed that their feet have developed webbing similar to that of ducks.*	*under 2cm*	*2cm*
How (method)	Tỉnh Bắc Ninh có ý nghĩa như thế nào với Hà Nội?	Tuy nhiên không vì không vì mà nhỏ quá mà Bắc Ninh lại kém phát triển của anh ạ Bắc Ninh được coi **là thành phố vệ tinh của Hà Nội** vì vậy là có rất nhiều những cái khu công nghiệp đem lại giá trị kinh tế cao cho Việt Nam giống như là Samsung	là thành phố vệ tinh	thành phố vệ tinh
	How important is Bac Ninh to Hanoi?	*The fact that Bac Ninh is small does not mean it is less developed, my friend. Bac Ninh which is considered* ***as a satellite city of Hanoi****, there are many industrial zones that bring high economic value to Vietnam, such as Samsung.*	*as a satellite city*	*a satellite city*
Why	Vì sao khó xác định được lượng nước chính xác để trộn bột?	sử dụng từ 200 cho tới 250 g nước trong các bạn thì à mà mình sử dụng nó sẽ **phụ thuộc vào cái bột hút nước nhiều hay ít** có nghĩa là nếu mà Bột mới thì nó sẽ hút nước ít hơn là bột củ và cái bột mới và một củ thì các bạn sẽ tính vào cái ngày sản xuất	phụ thuộc vào bột hút nước nhiều hay ít	bột hút nước nhiều hay ít
	Why is it hard to determine the exact amount of water to mix the dough?	*The amount of water you should use, between 200 and 250 grams, will* ***depend on the absorbency of the flour*** *which means The newer flour will be less absorbency than the old one. The old flour is determined by its production date.*	*depend on the absorbency of the flour*	*the absorbency of the flour*

	Question	**Transcript**	**Reference**	**Prediction**
Where	Món bánh hỏi nổi tiếng nhất ở đâu?	cái món bánh hỏi này nó có rất nhiều nơi nhưng để thành danh thì là **mảnh đất võ Bình Định nói chung và Quy Nhơn nói riêng** nếu như chúng ta mà đi về đây mà không thưởng thức món này thì có lẽ đó là một cái thiếu sót	Bình Định nói chung và Quy Nhơn nói riêng	mảnh đất võ Bình Định
	Where is the best place to try bánh hỏi?	*This dish, called ’bánh hỏi’, is available in many places, but to taste the best, one must visit **the land of martial arts, Binh Dinh in general and Quy Nhon city in particular** If we come here and do not try this dish, it would be a regrettable omission.*	*Binh Dinh in general and Quy Nhon city in particular*	*the land of martial arts, Binh Dinh*
When	Lúc nào thì có thể cho bánh vào dầu để chiên?	Bây giờ thì mình sẽ đem đi chim nha mình cho dầu ăn vào trong trở và động cơ nó nóng lên nhé em rửa và mình thấy **cái đầu nó sôi lăn tăn** đây nè đó Mình sẽ cho bánh vào nhá	đầu nó sôi lăn tăn	cái đầu nó sôi lăn tăn
	When cake should be put in the pan?	*I will fry the cake in a moment, now put the cooking oil to the pan and wait for it to be heated it up. Once it’s hot, test it by dipping a chopstick in; if there **bubbles form around the tip** it means the oil is ready. Then, I’ll add the cake to the pan.*	*bubbles form around the tip*	*bubbles form around the tip**
Who	Công thức được chia sẻ này phù hợp với những ai?	nên hôm nay là thay chia sẻ cái công thức bột này **tương đối dễ cho các bạn mới bắt đầu** do đó là nếu mà các bạn cảm thấy là cái nguyên liệu này nó khó tìm thì bạn có thể thay thế linh hoạt hơn thì vẫn cái bột vẫn chủ đạo nhất đó chính là một mì	các bạn mới bắt đầu	tương đối dễ cho các bạn mới bắt đầu
	Who does this recipe best suit?	*So today, Natha share a flour recipe, **relatively easy for beginners** If you find it difficult to find the original ingredients, you can still be flexible and replace them with other alternatives. The primary ingredient is still wheat flour.*	*beginners*	*relatively easy for beginners*

Table 11: Error examples for each question type of XLM-R model. The corresponding corrected English translations are also provided.