ViQA-COVID: COVID-19 Machine Reading Comprehension Dataset for Vietnamese

Anonymous ACL submission

Abstract

After two years of appearance, COVID-19 has negatively affected people and normal life around the world. As in November 2021, there are more than 250 million cases and five million deaths worldwide (including nearly one million cases and over twenty-two thousand deaths in Vietnam). Economy and society are both severely affected. The variant of COVID-19, Delta, has broken disease prevention measures of countries and rapidly increased number of infections. Resources overloading in treatment and epidemics prevention is happening all over the world. It can be seen that, application of artificial intelligence (AI) to support people at this time is extremely necessary. There have been many studies applying AI to prevent COVID-19 which are extremely useful, and studies on machine reading comprehension (MRC) are also in it. Realizing that, we created the first MRC dataset about COVID-19 for Vietnamese: ViQA-COVID and can be used to build models and systems, contributing to disease prevention. Besides, ViQA-COVID is also the first multi-span extraction MRC dataset for Vietnamese, we hope that it can contribute to promoting MRC studies in Vietnamese and multilingual.

1 Introduction

Delta - a so-far-most-dangerous variant of SARS-CoV-2 has shown its danger in recent months. Specifically, on average, each day there are around five hundreds thousands new cases and around ten thousands deaths worldwide. The uncontrollably rapid spread leads to the overwhelming of resources in disease prevention: medical staff, medical equipment manufacturing workers, data analysts, anti-epidemic support teams, etc. In the long run, this will have serious economic, social, as well as human impacts.

The application of intelligent and automated systems such as artificial intelligence and machine learning models to assist and replace work for humans is essential: to help reduce the work load on staffs and hence, can help the fight against the pandemic more effective. Many researches on artificial intelligence have been applied and brought positive effects, greatly benefitting society. Among them, natural language processing (NLP) systems have worked extremely well and there will be many more applications that can be deployed in the future. Specifically, NLP applications in COVID-19 prevention can be mentioned as: information extraction; patient and location information query; automatic documents summarization; automatic question and answer about the pandemic system; epidemic related sentiment analysis, etc.

In the above application systems, MRC systems are used in many applications, but most of them only work in high-resources languages such as English, Chinese. It can be seen that dataset plays a very important role in the development of NLP systems in particular and machine learning in general. Thus, low-resource languages enrichment is the first step to create a MRC system for these languages. With that in mind, we introduce ViQA-COVID, a multi-span extraction MRC dataset related to COVID-19 for Vietnamese with the desire to develop Vietnamese and multilingual MRC systems and help in fighting against COVID-19. Dataset comprises question-answer pairs based on CDC case reports, assigned to COVID-19 prevention teams and news on Vietnam’s reputation online newspapers.

In the next section, we review on the related works. Section 3 presents about datasets, statistics and annotation process. Section 4 is devoted for experiments set up. The results and benchmark are described in Section 5. Section 6 summarizes the study and presents further research directions.
2 Related Work

In recent years, COVID-19 has spurred research in many fields especially in AI related ones. In the field of computer vision, researchers (Wang et al., 2020a) designed COVID-Net to detect COVID-19 cases from chest X-ray (CXR) images and introduced COVIDx, a dataset consisting of 13,975 CXR images across 13,870 patient cases. In (Wang et al., 2020b), three masked face datasets: Masked Face Detection Dataset (MFDD), Real-world Masked Face Recognition Dataset (RMFRD), and Simulated Masked Face Recognition Dataset (SMFRD) that helped a lot in detecting and reminding people to wear masks (one of the most effective measures to prevent covid-19), are introduced. The image editing approach and datasets: Correctly Masked Face Dataset (CMFD), Incorrectly Masked Face Dataset (IMFD), as well as their combination - masked face detection (MaskedFace-Net) are introduced in (Cabani et al., 2020). MaskedFace-Net has been applied to detect whether people are wearing masks and wearing them correctly.

In the field of NLP, COVID-QA (Möller et al., 2020) is an MRC dataset consisting of 2,019 pairs of questions - answers labeled by experts, with data sources collected from CORD-19. COVID-QA is widely used in evaluating MRC tasks and applied to tasks related to COVID-19. CovidQA (Tang et al., 2020) is one of the first Question Answering datasets, consisting of pairs of questions - articles and answers that are articles related to the question. CovidDialog (Ju et al., 2020) provides a dataset including doctor-patient conversations (603 consultations and 1,232 utterances in English and 399 consultations and 8,440 utterances in Chinese). Using CovidDialog, researchers (Zeng et al., 2020) have developed a medical dialogue system to provide information related to the pandemic. Phoner_COVID (Truong et al., 2021) is a Vietnamese NER dataset about COVID-19 which defined 10 entities related to COVID-19 patients information. In addition, there are many research works that have been highly applicable and have greatly supported countries in preventing COVID-19.

With the rapid development of NLP in Vietnam, many new datasets have been introduced. From collecting 174 articles on the Vietnamese Wiki and through a five-phase annotate process, UIT-ViQuAD (Nguyen et al., 2020a) was created with more than 23,000 question-answer pairs based on 5,109 passages. UIT-ViQuAD is a single span-extraction MRC dataset widely used in span-extraction MRC task Vietnamese besides UIT-ViNewsQA (Nguyen et al., 2020b), a dataset in healthcare domain consisting of 22,057 question-answer pairs based on 4,416 articles health report. In addition, ViMMRC (Nguyen et al., 2020c) is a multiple-choice dataset and includes 2,783 multiple-choice questions based on 417 Vietnamese texts. With the task of sentence extraction-based MRC, UIT-ViWikiQA (Do et al., 2021) is the first Vietnamese sentence extraction-based MRC dataset, created from converting the UIT-ViQuAD dataset. UIT-ViWikiQA includes 23,074 question-answer pairs, based on 5,109 passages.

In addition to the studies on COVID-19 and MRC datasets for Vietnamese, we also consulted other famous MRC datasets such as: SQuAD1.1 (Rajpurkar et al., 2016), SQuAD2.0 (Rajpurkar et al., 2018), GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), MASH-QA (Zhu et al., 2020), QUOREF (Dasigi et al., 2019) and DROP (Dua et al., 2019).

The above studies helped us to complete our research.

3 Dataset

In this section, we will describe ViQA-COVID in detail, annotation processing and statistics about the dataset.

Being supplied data on cases, reflections and frequently asked people’s questions, from CDC Vietnam, we support them in data processing to make statistical reports, forecast epidemics, extract patient’s information and collect people’s feedback. This work is essential in helping the understand the situation and make decisions to prevent COVID-19. However, it is extremely complicated to process data on some hundreds thousands of cases as well as complaints, questions from people every day by hand. Unfixed-form data and complexity of Vietnamese make it difficult to handle with rule-based approach. With the development of deep learning, it can solve the above problems. Based on deep learning, a MRC system can correctly return information based on passages and questions. For example: From the patient’s epidemiological information, the medical team asks: “Who has the COVID-19 patient been in contact with?”. MRC system can answer correctly and the medical team can isolate and treat...
Passage:

**English:** On 12/8, 13/8, 14/8/2020, patient was in a room in the concentrated isolation area for medical staff. On 15/8/2020, patient was sampled for oropharyngeal fluid testing (3rd time), on 16/8/2020, patient had (+) result with SARS-CoV-2 virus. Currently, patient is isolated and treated at Da Nang Lung Hospital.

**Question:** Bệnh nhân có kết quả dương tính vào ngày nào? (What date did the patient test positive?)

**Answer:** 16/8/2020

Passage:
**Vietnamese:** Thong tin dinh te: khoảng 07 giờ 00 ngày 26/7/2020, bệnh nhân trở về nhà và tiếp xúc với những người trong gia đình. Khoảng 07 giờ 00 ngày 27/7/2020, bệnh nhân được cách ly tại Bệnh viện đến ngày 02/8/2020, sáng ngày 03/8/2020, tại Bệnh viện Đà Nẵng, bệnh nhân được lấy mẫu xét nghiệm dịch hầu họng (lần 2) và có kết quả (+) với vi rút SARS-CoV-2. Bệnh nhân ở cùng phòng với anh D.T (bảo vệ Bệnh viện Đà Nẵng).

**English:** Epidemiological information: around 7:00 am on 26/7/2020, patient returned home and contacted with family members. Around 7:00 am on 27/7/2020, patient was isolated at Hospital until 02/8/2020. On the morning of 03/8/2020, at Da Nang Hospital, patient was sampled oropharyngeal fluid testing (2nd time) and got a (+) result for SARS-CoV-2 virus. Patient was in the same room with D.T (security guard Da Nang Hospital).

**Question:** Bệnh nhân đã tiếp xúc với những ai? (Who has the patient been in contact with?)

**Answer:** những người trong gia đình, anh Đ.T (bảo vệ Bệnh viện Đà Nẵng) (family members, D.T (security guard Da Nang Hospital))

Figure 1: Examples include passage, question and answer from ViQA-COVID. Bold words in passage are answers.

---

3.1 Annotation

Annotation includes following phases:

- With limited time and resource, annotating all the data is not possible. Therefore, we reviewed and chose report cases that are as informative and structurally diverse as possible. Data was encrypted sensitive information, corrected typing and grammar errors. After data cleaning, we collected a total of 537 passages.

- Data is manually annotated. Question-answer pairs in ViQA-COVID are based on the information CDC needs to support patients and prevent diseases, as well as questions from people about the epidemic situation. For example: "What places have patients been to?", "Where are the epidemic locations that I need to be aware of?", etc. Annotators will read each passage, create questions and mark spans for corresponding answers (a answer can include multi-span). Questions are diversified and avoiding duplication. Question-answer pairs are cross-checked to eliminate errors.

- We have proceeded to collect more data from reputable online portals and online news sites to diversify dataset. This data is also reviewed, manually annotated, and cross-checked.

- We reviewed and cross-checked again to complete ViQA-COVID dataset.

3.2 Statistics

ViQA-COVID after completion has a total of 6,444 question-answer pairs based on 537 passages. To our knowledge, ViQA-COVID is the first multi-span extraction MRC dataset on COVID for Vietnamese. Details of the statistics are in Table 1.

It can be seen that, because ViQA-COVID is a...
domain-specific dataset (COVID-19 and Health), the vocab size is not too large. In addition, the percentage of multi-span answers is quite high compared to most multi-span MRC datasets, around 20%.

We statistic question types in the dataset as follows: What: 19.3%, How: 3.33%, How many: 10.2%, Where: 17.16%, When: 36.61%, Who: 8.38%, Others: 5.02%. Like many others languages, there may be some variations in Vietnamese each type of question. Statistical description of question words in ViQA-COVID is shown in Figure 2.

4 Experiments

In this section, we present experiments with the state-of-the-art MRC models on ViQA-COVID.

4.1 Models

Since BERT (Devlin et al., 2019) - a pretrained model using Transformer (Vaswani et al., 2017) architecture appeared in 2019, it has created a strong development in the field of natural language processing. State-of-the-art performance on NLP tasks increased rapidly thanks to improved models from both BERT and the Transformer architecture. It can be said that they are the two main factors that create a new era for NLP. In this experimental part, we used variants of BERT to evaluate on ViQA-COVID. These models have achieved state-of-the-art results on many MRC tasks.

- **mBERT**: twelve layers with twelve self-attention heads BERT is trained on multi-lingual datasets (including Vietnamese). Since its launch in 2019, mBERT has performed very well in multi-lingual MRC and NLP tasks.

- **XLM-R** (Conneau et al., 2020): based on RoBERTa (Liu et al., 2019) - an optimal BERT-based approach, XLM-R was trained on over two terabytes of cleaned CommonCrawl (Wenzek et al., 2019) data in 100 languages. XLM-R outperformed mBERT in many cross-lingual benchmarks and other tasks. We evaluated two model - XLM-R base: 12 layers with 8 self-attention heads and XLM-R large: 24 layers with 16 self-attention heads.

- **PhoBERT** (Nguyen and Nguyen, 2020): based on RoBERTa, PhoBERT is a Vietnamese model which improved the state-of-the-art many Vietnamese NLP tasks. PhoBERT is trained on over 20 gigabytes of word-level data (while other models train with syllable data). We also evaluated two models: PhoBERT base: 12 layers with 12 self-attention heads and PhoBERT large: 24 layers with 16 self-attention heads.

4.2 Input Processing

Statistics from Table 2 show that most passages are in excess of 512 tokens in length. Whereas maximum length of models’ input feature is 512 tokens. To deal with very long passage, we split one example into input features, each of the length

<table>
<thead>
<tr>
<th>Number of passages</th>
<th>Train</th>
<th>Dev.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>284</td>
<td>139</td>
<td>114</td>
</tr>
<tr>
<td>Number of questions</td>
<td>3408</td>
<td>1668</td>
<td>1368</td>
</tr>
<tr>
<td>Average passage length</td>
<td>336.8</td>
<td>269.1</td>
<td>252.7</td>
</tr>
<tr>
<td>Average question length</td>
<td>11.2</td>
<td>9.5</td>
<td>11.1</td>
</tr>
<tr>
<td>Passage vocabulary size</td>
<td>6659</td>
<td>3882</td>
<td>3089</td>
</tr>
<tr>
<td>Question vocabulary size</td>
<td>1071</td>
<td>606</td>
<td>601</td>
</tr>
<tr>
<td>Number of multi-span answers (%)</td>
<td>712 (20.9)</td>
<td>351 (21.0)</td>
<td>291 (21.3)</td>
</tr>
<tr>
<td>Number of single-span answers (%)</td>
<td>2288 (67.1)</td>
<td>1147 (68.8)</td>
<td>927 (67.8)</td>
</tr>
<tr>
<td>Number of non-span answer (%)</td>
<td>408 (12.0)</td>
<td>170 (10.2)</td>
<td>150 (10.9)</td>
</tr>
</tbody>
</table>

Table 1: ViQA-COVID overview

<table>
<thead>
<tr>
<th>Passage Length</th>
<th>Train</th>
<th>Dev.</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 128 tokens</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>128 - 256 tokens</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>256 - 384 tokens</td>
<td>25</td>
<td>11</td>
<td>8</td>
<td>44</td>
</tr>
<tr>
<td>384 - 512 tokens</td>
<td>38</td>
<td>20</td>
<td>18</td>
<td>76</td>
</tr>
<tr>
<td>≥ 512 tokens</td>
<td>260</td>
<td>119</td>
<td>96</td>
<td>475</td>
</tr>
</tbody>
</table>

Table 2: Passage length and Question length statistics
is shorter than model’s maximum length. In case the answer lies at the position that long passage was split, we create an overlap feature between two features (controlled by stride parameter).

PhoBERT is trained with both syllable-level and word-level tokens. Unlike English, words in Vietnamese can be compound words, i.e. one word with single meaning may be a combinations of two or more single words and in most of the cases, the meaning of the compound word is very different from their components. Thus, input sentences are segmented by word segmentation which can represent them in either syllable or word-level. Therefore, word segmentation joins syllables with a "_" sign to indicate it’s a word and makes sentences have clearer meanings. With that idea, PhoBERT outperformed XLM-R in many Vietnamese-specific NLP tasks. In our experiment, we use RDRSegmenter (Nguyen et al., 2018) from VnCoreNLP (Vu et al., 2018) as word and sentence segmentation.

### 4.3 Multi-span Approach

For the BERT-style models, we use sequence tagging approach (Segal et al., 2020) for multi-span questions. Instead of predicting start and end probabilities like single-span questions, we predict the tag for each token. The familiar tags used are B, I, O, where B is the starting token and I is the subsequent token in output span, O is the token that is not part of an output span. Multi-span can be extracted based on B, O tokens. Figure 3 illustrates this approach in detail.

### 4.4 Training

BERT-style models have maximum input features length of 384 (PhoBERT of 256) with stride parameter of 128. We fine-tuned models with AdamW (Loshchilov and Hutter, 2019), weight decay of 0.01, learning rate of 5e-5 and batch size of 32, in 30 training epochs on a NVIDIA Tesla P100 GPU via Google Colaboratory. Task performance was evaluated after each epoch on the development set.

### 5 Results

We evaluated models’ performance on ViQA-COVID using exact match (EM) and F1-score. Results are shown in Table 3. In which, XLM-R\_large outperforms other models with 83.37% F1-score and 68.82% EM on development set and 85.97% F1-score and 72.00% EM on test set. We also evaluated the performance of the models on single-span...
<table>
<thead>
<tr>
<th>Model</th>
<th>Dev.</th>
<th></th>
<th>Test</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-Span</td>
<td>Multi-Span</td>
<td>All</td>
<td>Single-Span</td>
</tr>
<tr>
<td></td>
<td>EM</td>
<td>F1</td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>mBERT</td>
<td>45.10</td>
<td>51.09</td>
<td>30.52</td>
<td>61.81</td>
</tr>
<tr>
<td>PhoBERTbase</td>
<td>61.86</td>
<td>72.73</td>
<td>30.59</td>
<td>54.12</td>
</tr>
<tr>
<td>PhoBERTlarge</td>
<td>62.13</td>
<td>72.48</td>
<td>32.74</td>
<td>56.70</td>
</tr>
<tr>
<td>XLM-Rbase</td>
<td>78.90</td>
<td>83.32</td>
<td>33.20</td>
<td>71.95</td>
</tr>
<tr>
<td>XLM-Rlarge</td>
<td>82.74</td>
<td>86.79</td>
<td>38.20</td>
<td>75.84</td>
</tr>
</tbody>
</table>

Table 3: Performances on development set and test set

<table>
<thead>
<tr>
<th>Question Types</th>
<th>Dev. errors</th>
<th>Test errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td>173</td>
<td>163</td>
</tr>
<tr>
<td>Where</td>
<td>98</td>
<td>84</td>
</tr>
<tr>
<td>Who</td>
<td>83</td>
<td>72</td>
</tr>
<tr>
<td>Others</td>
<td>135</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 4: Question type errors on development set and test set

and multi-span answers. The models are quite accurate in predicting single-span answers but still have difficulties with multi-span answers, especially in terms of exact matching. Overall, XLM-R\textsubscript{large} performed quite well and the difficulty of ViQA-COVID is not too hard when compare to other MRC datasets.

5.1 Error analyst

Through empirical analysis with the best model XLM-R\textsubscript{large}, we have counted the number of incorrect answers in the development set and test set. The development set has 489/1,668 incorrect answers of which 162 multi-span, 246 single-span and 81 non-span answers. The test set has 414/1,368 incorrect answers of which 141 multi-span, 203 single-span, and 70 non-span answers. We divide these errors into four groups:

- The first group consists of answers that have the correct number of spans but have an excess or lack of words. The cases are mostly long addresses or time periods (e.g. “20/5/2020 to 30/5/2020” but the model can only predict “20/5” or “30/5”). These are also common mistakes in sequence tagging models.

- The second group includes answers that have an excess or lack of span. Mainly occurs when encountering questions about many places or about many people. For example: answering a question that lists people who have been in contact with the patient but also lists those who have not.

- The third group are completely incorrect answers (answers that have no correct span), often occurring in passages having a lot of noise. For example: Patient’s epidemiological report contains multiple dates, including dates of admission. When answering the question about the date of admission for COVID-19 infection, the model easily mistakenly answered to the date the patient was hospitalized for another illness because of the same keyword “admission”.

- The fourth group includes incorrect answers on other types of questions.

The statistics of the incorrect answers are shown in Table 4.

6 Conclusion

In this study, we introduced ViQA-COVID, the first multi-span MRC dataset about COVID-19 for Vietnamese. Our dataset consists of 6,444 question-answer pairs based on 537 passages related to COVID-19. We also experimented with different the state-of-the-art MRC models on ViQA-COVID. The results show that, XLM-R\textsubscript{large} outperforms other models with 83.37% F1-score and 68.82% EM on development set and 85.97% F1-score and 72.00% EM on test set. We hope that our dataset will contribute to the prevention of COVID-19 as well as the development of NLP for Vietnamese and multilingual.

References


