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 January 2022, there are more than 317 million cases and five million deaths worldwide (including nearly two million cases and over thirty-four thousand deaths in Vietnam). Economy and society are both severely affected. The variant of COVID-19, Omicron, 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. We will publicly release ViQA-COVID soon.

1 Introduction

Omicron - a 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.

In Vietnam, the number of cases is increasing very quickly. The information of the cases must be updated continuously to support the medical team to capture information and promptly treat the patient. On national portals, important information about the epidemic such as the number of cases, time and location related to the epidemic, people in contact with the patient, also needs to be updated quickly so that people can grasp the information and protect themselves. In addition, the hotlines and portals receive a lot of questions and reflections from the people every day. It can be seen that the amount of data generated daily is very large and difficult to handle manually. Thus, a system to extract information and answer questions like the machine reading comprehension (MRC) system is extremely necessary at the present time. It will be an aid for the prevention of COVID-19 or even other diseases in the future.

To build a COVID-19 MRC system, a COVID-19 MRC dataset is required. As a matter of fact, sufficient MRC dataset on COVID-19 for Vietnamese has yet to be released. Therefore, we created ViQA-COVID, a multi-span extraction MRC dataset about COVID-19 for Vietnamese based on official data from Centers for Disease Control and Prevention (CDC) Vietnam and reputable online news sites. In addition, ViQA-COVID is also the first multi-span extraction MRC dataset for Vietnamese. The goal of this research is to contribute to building data sources for low-resource languages like Vietnamese.

In the next section, related works will be covered. 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 (RM-FRD), 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. (Zhang et al., 2021) publicly released COUGH, a COVID-19 FAQ dataset includes 15,919 FAQ items, 1,236 human-paraphrased user queries and each query has 32 human-annotated FAQ items. 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 datasets 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, ViQA-COVID, annotation processing and statistics about the dataset is described in detail.

CDC daily receives a large amount of data on cases, reflections and questions from people. Extracting and processing information from this data source is essential to helping medical teams understand the situation and make decisions to prevent COVID-19. However, handling huge amounts of data by hand is extremely complex. In addition, unfixed-form data and complexity of Vietnamese make it difficult to handle with rule-based approach. Based on previous studies as (Nguyen et al., 2020a), (Zhu et al., 2020), (Segal et al., 2020), etc., it can be seen that a MRC system based on deep learning can solve the above problems. 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 those people quickly. In addition, MRC system can help answer people's questions about disease, policies, ways to prevent COVID-19

Passage:
Vietnamese: Thông tin dịch tễ: 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 Đ.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 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: An example include passage, question and answer from ViQA-COVID. Bold words in passage are answers

and so on. To be able to achieve the aforementioned purposes, MRC system needs to train with MRC datasets. Therefore, ViQA-COVID has been created as training data for such system. Figure 1 shows an example from ViQA-COVID.

3.1 Annotation

The annotation team consists of five data analyst from CDC annotating and reviewing data, and three experts from CDC advising on the questions and information to annotate on the data. In general, the annotation process includes following phases:

- **Collect and clean passage data from CDC:** With limited time and resource, annotating all the data is not possible. Therefore, report cases are chosen on the basis of informativeness and structural diversity. Data was encrypted sensitive information (so as not to violate privacy issues), corrected typing and grammar errors. After data cleaning, a total of 537 passages were collected.

- **Create and cross-check question-answer pairs:** 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 avoid duplication. Question-answer pairs are cross-checked to eliminate errors.

- **Collect data from other sources, annotate and cross-check:** More data from reputable online portals and online news sites were collected to diversify dataset. This data is also reviewed, manually annotated and cross-checked.

- **Review and recheck:** To ensure data was clean and did not violate privacy issues, 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%.

Question types in the dataset is distributed 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, each type of question may be expressed in numerous ways. Statistical description of question words in ViQA-COVID is shown in Table 2.

4 Experiments

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

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

Table 1: ViQA-COVID overview

Question Types	Question Words
What (19.3%)	là gì (10.5%)
Where (17.2%)	đâu (7.3%)
When (36.6%)	ngày nào (10.4%)
Who (8.4%)	ai (6.2%)
How (3.3%)	thế nào (1.1%)
How many (10.2%)	bao nhiêu (9.6%)

Table 2: Question types and questions words distribution in ViQA-COVID

Passage Length	Train	Dev.	Test	Total
< 128 tokens	0	0	2	2
128 - 256 tokens	12	1	2	15
256 - 384 tokens	25	11	8	44
384 - 512 tokens	38	20	18	76
$\geq$ 512 tokens	260	119	96	475
	335	151	126	612

Table 3: Passage length statistics

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 Common-Crawl (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$_{\text{base}}$: 12 layers with 8 self-attention heads and XLM-R$_{\text{large}}$: 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$_{\text{base}}$: 12 layers with 12 self-attention heads and PhoBERT$_{\text{large}}$: 24 layers with 16 self-attention heads

4.2 Input Processing

Statistics from Table 3 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 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 Viet-

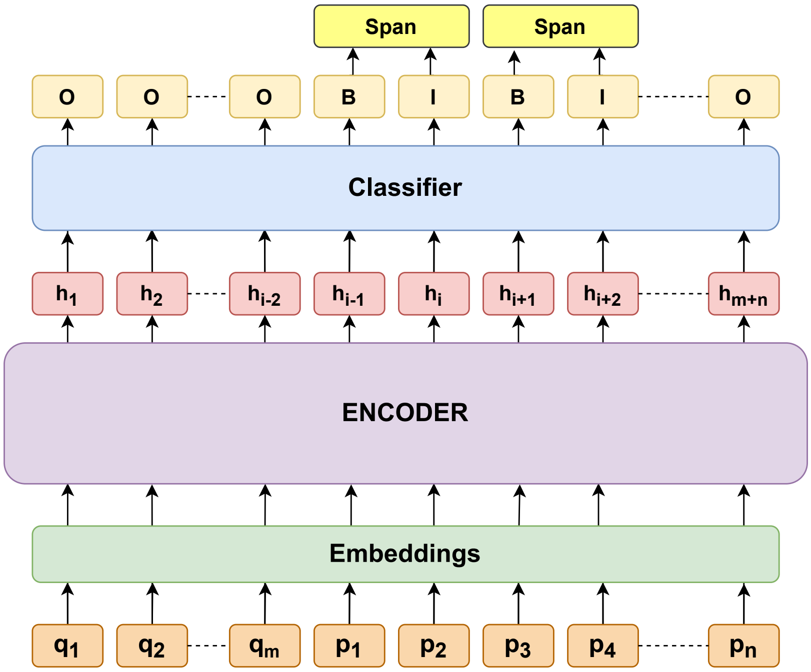

Figure 2: Illustrating the sequence tagging approach for multi-span questions. In which, $\{q_j\}_{j=1}^{m}$ are question tokens, $\{p_k\}_{k=1}^{n}$ are passage tokens and $\{h_i\}_{i=1}^{m+n}$ are the contextualized representations of the input tokens.

namese 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 2 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 4. In which, XLM-R$_{\text{large}}$ 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 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$_{\text{large}}$ performed quite well and the difficulty of ViQA-COVID is not too hard when compare to

Model	Dev.						Test					
	Single-Span		Multi-Span		All		Single-Span		Multi-Span		All	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
mBERT	45.10	51.09	30.52	61.81	40.83	54.44	46.28	51.14	37.64	65.98	43.49	55.96
PhoBERT$_{base}$	61.86	72.73	30.59	54.12	51.37	66.49	54.90	74.39	34.99	60.87	54.89	70.01
PhoBERT$_{large}$	62.13	72.48	32.74	56.70	52.28	67.19	64.65	74.21	37.25	62.14	55.77	70.30
XLM-R$_{base}$	78.90	83.32	33.20	71.95	64.62	79.77	81.23	85.13	41.27	77.83	68.34	82.78
XLM-R$_{large}$	**82.74**	**86.79**	**38.20**	**75.84**	**68.82**	**83.37**	**85.11**	**89.24**	**44.44**	**79.10**	**72.00**	**85.97**

Table 4: Performances on development set and test set

Question Types	Dev. errors	Test errors
When	173	163
Where	98	84
Who	83	72
Others	135	95

Table 5: Question type errors on development set and test set

other MRC datasets.

5.1 Error analyst

Through empirical analysis with the best model XLM-R$_{large}$, 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 5.

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$_{large}$ 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

Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi. 2020. Maskedface-net – a dataset of correctly/incorrectly masked face images in the context of covid-19. *Smart Health*.

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.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential 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 5925–5932, 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. *CoRR*, abs/2105.09043.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, and Pengtao Xie. 2020. Coviddialog: Medical dialogue datasets about covid-19. *https://github.com/UCSD-AI4H/COVID-Dialogue*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

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.

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras, and Mark Johnson. 2018. A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2582–2587.

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, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. New vietnamese corpus for machine readingcomprehension of health news articles. *CoRR*, abs/2006.11138.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. 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.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080, Online. Association for Computational Linguistics.

Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly bootstrapping a question answering dataset for COVID-19. *CoRR*, abs/2004.11339.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. 2020a. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549.

Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, Heling Chen, Yu Miao, Zhibing Huang, and Jinbi Liang. 2020b. Masked face recognition dataset and application. *CoRR*, abs/2003.09093.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data.

Guangtao Zeng, Qingyang Wu, Yichen Zhang, Zhou Yu, Eric Xing, and Pengtao Xie. 2020. Develop medical dialogue systems for covid-19. *https://github.com/UCSD-AI4H/COVID-Dialogue*.

Xinliang Frederick Zhang, Heming Sun, Xiang Yue, Simon Lin, and Huan Sun. 2021. COUGH: A challenge dataset and models for COVID-19 FAQ retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 3759–3769.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.