COVID-19 Named Entity Recognition for Vietnamese

Thinh Hung Truong, Mai Hoang Dao and **Dat Quoc Nguyen**
VinAI Research, Hanoi, Vietnam
{v.thinhth88, v.maidh3, v.datnq9}@vinai.io

Abstract

The current COVID-19 pandemic has lead to the creation of many corpora that facilitate NLP research and downstream applications to help fight the pandemic. However, most of these corpora are exclusively for English. As the pandemic is a global problem, it is worth creating COVID-19 related datasets for languages other than English. In this paper, we present the first manually-annotated COVID-19 domain-specific dataset for Vietnamese. Particularly, our dataset is annotated for the named entity recognition (NER) task with newly-defined entity types that can be used in other future epidemics. Our dataset also contains the largest number of entities compared to existing Vietnamese NER datasets. We empirically conduct experiments using strong baselines on our dataset, and find that: automatic Vietnamese word segmentation helps improve the NER results and the highest performances are obtained by fine-tuning pre-trained language models where the monolingual model PhoBERT for Vietnamese (Nguyen and Nguyen, 2020) produces higher results than the multilingual model XLM-R (Conneau et al., 2020). We publicly release our dataset at: https://github.com/VinAIResearch/PhoNER_COVID19.

1 Introduction

As of early November 2020, the total number of COVID-19 cases worldwide has surpassed 50M.[1] The world is once again hit by a new wave of COVID-19 infection with record-breaking numbers of new cases reported everyday. Along with the outbreak of the pandemic, information about the COVID-19 is aggregated rapidly through different types of texts in different languages (Aizawa et al., 2020). Particularly, in Vietnam, text reports containing official information from the government about COVID-19 cases are presented in great detail, including de-identified personal information, travel history, as well as information of people who come into contact with the cases. The reports are frequently kept up to date at reputable online news sources, playing a significant role to help the country combat the pandemic. It is thus essential to build systems to retrieve and condense information from those official sources so that related people and organizations can promptly grasp the key information for epidemic prevention tasks, and the systems should also be able to adapt and sync quickly with epidemics that take place in the future. One of the first steps to develop such systems is to recognize relevant named entities mentioned in the texts, which is also known as the NER task.

Compared to other languages, data resources for the Vietnamese NER task are limited, including only two public datasets from the VLSP 2016 and 2018 NER shared tasks (Huyen and Luong, 2016; Nguyen et al., 2018b). Here, the VLSP-2018 NER dataset is an extension of the VLSP-2016 NER dataset with more data. These two datasets only focus on recognizing generic entities of person names, organizations, and locations in online news articles. Thus, making them difficult to adapt to the context of extracting key entity information related to COVID-19 patients. This leads to our work's main goals that are: (i) To develop a NER task in the COVID-19 specified domain, that potentially impacts research and downstream applications, and (ii) To provide the research community with a new dataset for recognizing COVID-19 related named entities in Vietnamese.

In this paper, we present a named entity annotated dataset with newly-defined entity types that can be applied to future epidemics. The dataset contains informative sentences related to COVID-19, extracted from articles crawled from reputable Vietnamese online news sites. Here, we do not consider other types of popular social media in Vietnam such as Facebook as they contain much

[1] https://www.worldometers.info/coronavirus/worldwide-graphs/#total-cases

Label	Definition
PATIENT_ID	Unique identifier of a COVID-19 patient in Vietnam. An PATIENT_ID annotation over "X" refers to as the X^{th} patient having COVID-19 in Vietnam.
PERSON_NAME	Name of a patient or person who comes into contact with a patient.
AGE	Age of a patient or person who comes into contact with a patient.
GENDER	Gender of a patient or person who comes into contact with a patient.
OCCUPATION	Job of a patient or person who comes into contact with a patient.
LOCATION	Locations/places that a patient was presented at.
ORGANIZATION	Organizations related to a patient, e.g. company, government organization, and the like, with structures and their own functions.
SYMPTOM&DISEASE	Symptoms that a patient experiences, and diseases that a patient had prior to COVID-19 or complications that usually appear in death reports.
TRANSPORTATION	Means of transportation that a patient used. Here, we only tag the specific identifier of vehicles, e.g. flight numbers and bus/car plates.
DATE	Any date that appears in the sentence.

Table 1: Definitions of entity types in our annotation guidelines. We do not annotate nested entities.

noisy information and are not as reliable as official news sources. We then empirically evaluate strong baseline models on our dataset. Our contributions are summarized as follows:

- We introduce the first manually annotated Vietnamese dataset in the COVID-19 domain. Our dataset is annotated with 10 different named entity types related to COVID-19 patients in Vietnam. Compared to the VLSP-2016 and VLSP-2018 Vietnamese NER datasets, our dataset has the largest number of entities, consisting of 35K entities over 10K sentences.
- We empirically investigate strong baselines on our dataset, including BiLSTM-CNN-CRF (Ma and Hovy, 2016) and the pre-trained language models XLM-R (Conneau et al., 2020) and PhoBERT (Nguyen and Nguyen, 2020). We find that: (i) Automatic Vietnamese word segmentation helps improve the NER results, and (ii) The highest results are obtained by fine-tuning the pre-trained language models, where PhoBERT does better than XLM-R.
- We publicly release our dataset for research or educational purposes. We hope that our dataset can serve as a starting point for future COVID-19 related Vietnamese NLP research and applications.

2 Related work

Most COVID-19 related datasets are constructed from two types of sources. The first one is scientific publications, including the datasets CORD-19 (Wang et al., 2020) and LitCovid (Chen et al., 2020), that help facilitate many types of research works, such as building search engines to retrieve relevant information from scholarly articles (Esteva et al., 2020; Zhang et al., 2020; Verspoor et al., 2020), question answering and summarization (Lee et al., 2020; Su et al., 2020). Recently, Colic et al. (2020) fine-tune a BERT-based NER model on the CRAFT corpus (Verspoor et al., 2012) to recognize and then normalize biomedical ontology and terminology entities in LitCovid.

The second type is social media data, particularly Tweets. COVID-19 related Tweet datasets are built for many analytic tasks such as identification of informative Tweets (Nguyen et al., 2020b), and disinformation detection and fact-checking (Shahi and Nandini, 2020; Alam et al., 2020; Alsudias and Rayson, 2020). The most relevant work to ours is proposed by Zong et al. (2020), that aims to extract COVID-19 events reporting test results, death cases, cures and prevention from English Tweets. As Twitter is rarely used by Vietnamese people, we could not use it for data collection.

3 Our dataset

3.1 Entity types

We define 10 entity types with the aim of extracting key information related to COVID-19 patients, which are especially useful in downstream applications. In general, these entity types can be used in the context of not only the COVID-19 pandemic but also in other future epidemics. The description of each entity type is briefly described in Table 1. See the Appendix for entity examples as well as some notices over the entity types.

3.2 COVID-19 related data collection

We first crawl articles tagged with "COVID-19" or "COVID" keywords from the reputable Vietnamese online news sites, including VnExpress,[2] ZingNews,[3] BaoMoi[4] and ThanhNien.[5] These articles are dated between February 2020 and August 2020. We then segment the crawled news articles' primary text content into sentences using RDRSegmenter (Nguyen et al., 2018a) from VnCoreNLP (Vu et al., 2018).

To retrieve informative sentences about COVID-19 patients, we employ BM25Plus (Trotman et al., 2014) with search queries of common keywords appearing in sentences that report confirmed, suspected, recovered, or death cases as well as the travel history or location of the cases. From the top 15K sentences ranked by BM25Plus, we manually filter out sentences that do not contain information related to patients in Vietnam, thus resulting in a dataset of 10027 raw sentences.

3.3 Annotation process

We develop an initial version of our annotation guidelines and then randomly sample a pilot set of 1K sentences from the dataset of 10027 raw sentences for the first phase of annotation. Two of the guideline developers are employed to annotate the pilot set independently. Following Brandsen et al. (2020), we utilize F_1 score to measure the inter-annotator agreement between the two annotators at the entity span level, resulting in an F1 score of 0.88. We then host a discussion session to resolve annotation conflicts, identify complex cases, and refine the guidelines.

In the second annotation phase, we divide the whole dataset of 10027 sentences into 10 non-overlapping and equal subsets. Each subset contains 100 sentences from the pilot set from the first annotation phase. For this second phase, we employ 10 annotators who are undergraduate students with strong linguistic abilities (here, each annotator annotates a subset, paid 0.05 USD per sentence). Annotation quality of each annotator is measured by F_1 calculated over the 100 sentences that already have gold annotations from the pilot set. All annotators are asked to revise their annotations until they achieve an F_1 of at least 0.92. Finally, we revisit each annotated sentence to make further corrections if needed, resulting in a final gold dataset of 10027 annotated sentences.

Note that when written in Vietnamese texts, in addition to marking word boundaries, white space is also used to separate syllables that constitute words. Therefore, the annotation process is performed at syllable-level text for convenience. To obtain a word-level variant of the dataset, we apply the RDRSegmenter to perform automatic Vietnamese word segmentation, e.g. a 4-syllable written text "bệnh viện Đà Nẵng" (Da Nang hospital) is word-segmented into a 2-word text "bệnh_viện$_{\text{hospital}}$ Đà_Nẵng$_{\text{Da_Nang}}$". Here, automatic Vietnamese word segmentation outputs do not affect gold boundaries of entity mentions.

Entity Type	Train	Valid.	Test	All
PATIENT_ID	3240	1276	2005	6521
PERSON_NAME	349	188	318	855
AGE	682	361	582	1625
GENDER	542	277	462	1281
OCCUPATION	205	132	173	510
LOCATION	5398	2737	4441	12576
ORGANIZATION	1137	551	771	2459
SYMPTOM&DISEASE	1439	766	1136	3341
TRANSPORTATION	226	87	193	506
DATE	2549	1103	1654	5306
# Entities in total	15767	7478	11735	34984
# Sentences in total	5027	2000	3000	10027

Table 2: Statistics of our dataset.

3.4 Data partitions

We randomly split the gold annotated dataset of 10027 sentences into training/validation/test sets with a ratio of 5/2/3, ensuring comparable distributions of entity types across these three sets. Statistics of our dataset is presented in Table 2.

4 Experiments

4.1 Experimental setup

We formulate the COVID-19 NER task for Vietnamese as a sequence labeling problem with the BIO tagging scheme. We conduct experiments on our dataset using strong baselines to investigate: (i) the influence of automatic Vietnamese word segmentation (here, input sentence can be represented in either syllable or word level), and (ii) the usefulness of pre-trained language models. The baselines include: BiLSTM-CNN-CRF (Ma and Hovy, 2016) and the pre-trained language models XLM-R (Conneau et al., 2020) and PhoBERT (Nguyen and

[2] https://vnexpress.net
[3] https://zingnews.vn
[4] https://baomoi.com
[5] https://thanhnien.vn

	Model	PAT.	PER.	AGE	GEN.	OCC.	LOC.	ORG.	SYM.	TRA.	DAT.	Mic-F_1	Mac-F_1
Syllable	BiL-CRF	0.953	0.855	0.943	0.947	0.588	0.915	0.808	0.801	0.794	0.976	0.906	0.858
	XLM-R_{base}	0.978	0.902	0.957	0.842	0.560	0.941	0.842	0.858	0.924	0.982	0.925	0.879
	XLM-R_{large}	**0.982**	0.933	0.962	0.958	0.692	**0.943**	0.853	0.854	0.943	0.987	0.938	0.911
Word	BiL-CRF	0.953	0.874	0.950	0.947	0.605	0.911	0.831	0.799	0.902	0.976	0.910	0.875
	PhoBERT$_{base}$	0.981	0.903	0.962	0.954	0.749	**0.943**	0.870	0.883	0.966	0.987	0.942	0.920
	PhoBERT$_{large}$	0.980	**0.944**	**0.967**	**0.968**	**0.791**	0.940	**0.876**	**0.885**	**0.967**	**0.989**	**0.945**	**0.931**

Table 4: Strict F_1 score for each entity type (denoted by its first 3 characters), and Micro- and Macro-average F_1 scores (denoted by Mic-F_1 and Mac-F_1, respectively). BiL-CRF abbreviates the baseline BiLSTM-CNN-CRF. **Syllable** and **Word** denote results obtained when using syllable- and word-level based dataset settings, respectively.

Hyper-parameter	**Value**
Optimizer	Adam
Learning rate	0.001
Mini-batch size	36
LSTM hidden state size	200
Number of BiLSTM layers	2
Dropout	[0.25, 0.25]
Character embedding size	50
Filter length, i.e. window size	3
Number of filters	30

Table 3: Hyper-parameters for BiLSTM-CNN-CRF.

Nguyen, 2020). XLM-R is a multi-lingual variant of RoBERTa (Liu et al., 2019), pre-trained on a 2.5TB multilingual dataset that contains 137GB of syllable-level Vietnamese texts. PhoBERT is a monolingual variant of RoBERTa, pre-trained on a 20GB word-level Vietnamese dataset.

We employ the BiLSTM-CNN-CRF implementation from AllenNLP (Gardner et al., 2018). Training BiLSTM-CNN-CRF requires input pre-trained syllable- and word-level embeddings for the syllable- and word-level settings, respectively. Thus we employ the pre-trained Word2Vec syllable and word embeddings for Vietnamese from Nguyen et al. (2020a). These embeddings are fixed during training. Optimal hyper-parameters that we grid-searched for BiLSTM-CNN-CRF are presented in Table 3. We utilize the *transformers* library (Wolf et al., 2020) to fine-tune XLM-R and PhoBERT for the syllable- and word-level settings, respectively, using Adam (Kingma and Ba, 2014) with a fixed learning rate of 5.e-5 and a batch size of 32 (Liu et al., 2019).

The baselines are trained/fine-tuned for 30 epochs. We evaluate the Micro-average F_1 score after each epoch on the validation set (here, we apply early stopping if we find no performance improvement after 5 continuous epochs). We then choose the best model checkpoint to report the final score on the test set. Note that each F_1 score reported is an average over 5 runs with different random seeds.

4.2 Main results

Table 4 shows the final entity-level NER results of the baselines on the test set. In addition to the standard Micro-average F_1 score, we also report the Macro-average F_1 score.

We categorize the results under two comparable settings of using syllable-level dataset and its automatically-segmented word-level variant for training and evaluation. We find that the performances of word-level models are higher than their syllable-level counterparts, showing that automatic Vietnamese word segmentation helps improve NER, e.g. BiLSTM-CNN-CRF improves from 0.906 to 0.910 Micro-F_1 and from 0.858 to 0.875 Macro-F_1.

We also find that fine-tuning the pre-trained language models XLM-R and PhoBERT helps produce better performances than BiLSTM-CNN-CRF. Here, PhoBERT outperforms XLM-R (Micro-F_1: 0.945 vs. 0.938; Macro-F_1: 0.931 vs. 0.911), thus reconfirming the effectiveness of pre-trained monolingual language models on the language-specific downstream tasks (Nguyen and Nguyen, 2020).

4.3 Error analysis

We perform an error analysis using the best performing model PhoBERT$_{large}$ that produces 353 incorrect predictions in total on the validation set.

The first error group consists of 69/353 instances with correct entity boundaries (i.e. exact spans) and incorrect entity labels. It is largely due to the fact that the model could not differentiate between LOCATION and ORGANIZATION entities. This is not surprising because of the ambiguity between these two entity types, in which the same entity mention may act as either LOCATION or ORGA-

NIZATION depending on the sentence context. Also, in terms of contact tracing, it would be more useful to label an organization-like entity mention as LOCATION if we can infer that a patient presented at that organization; however, such inference requires additional world knowledge about the entity. In addition, in this error group, the model also struggles to recognize OCCUPATION entities correctly. Recall that OCCUPATION entity mention must represent the job of a particular person labeled with PERSON_NAME or PATIENT_ID. Therefore, it may cause confusion to the model for deciding whether an occupation is linked to a determined person or not in a single sentence context.

The second error group contains 65/353 instances with inexact spans overlapped with gold spans but having correct entity labels. These errors generally happen with multi-word ORGANIZATION entity mentions, where (i) an ORGANIZATION entity contains a nested location inside its span, e.g. "*Bệnh viện Lao và Bệnh phổi Cần Thơ*" (Can Tho hospital for Tuberculosis and Lung disease; here, "Can Tho" is a province in Vietnam), or (ii) an organization is a subdivision of a larger organization, e.g. "*Khoa tim mạch - Bệnh viện Bạch Mai*" (Department of Cardiology - Bach Mai Hospital).[6]

The third group of 8/353 errors with overlapped inexact spans and incorrect entity labels does not provide us with any useful insight. The final group of remaining 211/353 errors is accounted for predicted entities corresponding with gold O labels. Particularly in the case of LOCATION, where generic mentions, such as "*Bệnh viện tỉnh*" (province hospital), "*Trạm y tế xã*" (commune medical station), "*chung cư*" (apartment), are recognized as entities, while in fact, they are not.

5 Conclusion

In this paper, we have presented the first manually-annotated Vietnamese dataset in the COVID-19 domain, focusing on the named entity recognition task. We empirically conduct experiments on our dataset to compare strong baselines and find that the input representations and the pre-trained language models all have influences on this COVID-19 related NER task. We hope that our dataset can serve as the starting point for further Vietnamese NLP research and applications in fighting the COVID-19 and other future epidemics.

[6]Word segmentation is not shown for simplification.

References

Akiko Aizawa, Frederic Bergeron, Junjie Chen, Fei Cheng, Katsuhiko Hayashi, Kentaro Inui, Hiroyoshi Ito, Daisuke Kawahara, Masaru Kitsuregawa, Hirokazu Kiyomaru, Masaki Kobayashi, Takashi Kodama, Sadao Kurohashi, Qianying Liu, Masaki Matsubara, Yusuke Miyao, Atsuyuki Morishima, Yugo Murawaki, Kazumasa Omura, Haiyue Song, Eiichiro Sumita, Shinji Suzuki, Ribeka Tanaka, Yu Tanaka, Masashi Toyoda, Nobuhiro Ueda, Honai Ueoka, Masao Utiyama, and Ying Zhong. 2020. A System for Worldwide COVID-19 Information Aggregation. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, and Preslav Nakov. 2020. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. *arXiv preprint*, arXiv:2005.00033.

Lama Alsudias and Paul Rayson. 2020. COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4573–4577.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2020. Keep up with the latest coronavirus research. *Nature*, 579:193.

Nico Colic, Lenz Furrer, and Fabio Rinaldi. 2020. Annotating the Pandemic: Named Entity Recognition and Normalisation in COVID-19 Literature. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

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.

Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. 2020. CO-Search: COVID-19 Information Retrieval with Semantic Search, Question Answering, and Abstractive Summarization. *arXiv preprint*, arXiv:2006.09595.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software*, pages 1–6.

Nguyen Thi Minh Huyen and Vu Xuan Luong. 2016. VLSP 2016 shared task: Named entity recognition. *Proceedings of Vietnamese Speech and Language Processing*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint*, arXiv:1412.6980.

Jinhyuk Lee, Sean S. Yi, Minbyul Jeong, Mujeen Sung, WonJin Yoon, Yonghwa Choi, Miyoung Ko, and Jaewoo Kang. 2020. Answering Questions on COVID-19 in Real-Time. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

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. *arXiv preprint*, arXiv:1907.11692.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020a. A pilot study of text-to-SQL semantic parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085.

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. 2018a. A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 2582–2587.

Dat Quoc Nguyen, Thanh Vu, Afshin Rahimi, Mai Hoang Dao, Linh The Nguyen, and Long Doan. 2020b. WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets. In *Proceedings of the 6th Workshop on Noisy User-generated Text*, pages 314–318.

Huyen TM Nguyen, Quyen T Ngo, Luong X Vu, Vu M Tran, and Hien TT Nguyen. 2018b. VLSP shared task: Named entity recognition. *Journal of Computer Science and Cybernetics*, 34(4):283–294.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – A Multilingual Cross-domain Fact Check News Dataset for COVID-19. In *Proceedings of the International Workshop on Cyber Social Threats*.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. 2020. CAiRE-COVID: A Question Answering and Query-focused Multi-Document Summarization System for COVID-19 Scholarly Information Management. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65.

Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher S. Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A. Baumgartner Jr., Michael Bada, Martha Palmer, and Lawrence E. Hunter. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinform.*, 13:207.

Karin Verspoor, Simon Šuster, Yulia Otmakhova, Shevon Mendis, Zenan Zhai, Biaoyan Fang, Jey Han Lau, Timothy Baldwin, Antonio Jimeno Yepes, and David Martinez. 2020. COVID-SEE: Scientific Evidence Explorer for COVID-19 related research. *arXiv preprint arXiv:2008.07880*.

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.

Lucy Lu Wang, Kyle Lo, et al. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Thomas Wolf, Lysandre Debut, et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Edwin Zhang, Nikhil Gupta, Raphael Tang, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, et al. 2020. Covidex: Neural ranking models and keyword search infrastructure for the covid-19 open research dataset. *arXiv preprint*, arXiv:2007.07846.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2020. Extracting COVID-19 Events from Twitter. *arXiv preprint*, arXiv:2006.02567.

Appendix

Annotation examples

Example 1:

Bệnh nhân " PAT 669 " là OCC bác sĩ làm việc tại LOC Bệnh viện Đa khoa Đồng Nai

Patient " PAT 669 " is a OCC doctor working at LOC Dong Nai General Hospital

Example 2:

ORG Bệnh viện Bệnh Nhiệt đới TP HCM xét nghiệm dương tính lần một đêm DAT 12/3 .

ORG Ho Chi Minh City Hospital for Tropical Diseases returns a positive test result in the evening of DAT 12/3 .

Example 3:

Hai nữ điều dưỡng LOC Bệnh viện Bạch Mai lây từ bên ngoài và lây nhiễm cho nhau.

Two nurses of LOC Bach Mai Hospital got infected from external source and then infected each other.

Example 4:

Bệnh nhân tử vong tại LOC Bệnh viện Phổi Đà Nẵng với chẩn đoán SYM viêm phổi nặng , SYM suy đa tạng không hồi phục , trên bệnh nhân SYM suy thận mạn giai đoạn cuối .

The patient died at LOC Da Nang Lung Hospital , diagnosed with SYM severe pneumonia with history of SYM unrecoverable multiorgan dysfunction syndrome , SYM terminal chronic kidney failure .

Here, PAT, OCC, LOC, DAT and SYM abbreviate PATIENT_ID, OCCUPATION, LOCATION, DATE and SYMPTOM&DISEASE, respectively. Recall that an annotation PATIENT_ID over "X" refers to as the Xth patient having COVID-19 in Vietnam (e.g. in Example 1: "669" refers to as the 669th patient).

Notices over entity types

We have two principles for selecting the ten entity types: (i) Entities should contain key information related to the COVID-19 patients (here, the information should be helpful in the context of contact tracing and monitoring the growth of the pandemic); and (ii) The availability of entity types in the text, i.e., how frequent does each of the entity types appear. This is decided based on manual observations of news articles.

In the context of contact tracing, it is more useful to broaden the scope of location. For example, when a patient is presented at an organization, we refer to that organization as a location if we can infer its specific location on the map. In Example 1, we would label the entity mention "*Bệnh viện Đa khoa Đồng Nai*" (Dong Nai General Hospital) with LOCATION as its provide information about the place that a patient used to be at. On the other hand, in Example 2, the entity mention "*Bệnh viện Bệnh Nhiệt đới TP HCM*" (Ho Chi Minh City Hospital for Tropical Diseases) is labeled as ORGANIZATION because it acts as the subject executing a specific action (i.e. reporting a test result).

For OCCUPATION, AGE and GENDER entities, we only tag them if we can link the corresponding entity mentions to a specific entity with NAME or PATIENT_ID label within the same sentence. In Example 1, "*bác sĩ*" (doctor) is the occupation of patient "*669*", thus we label this mention as an entity of type OCCUPATION. However, in Example 3, we do not label " *điều dưỡng*" (nurses) as OCCUPATION as we cannot link this mention to any determined person.

For SYMPTOM&DISEASE entities, we prefer the entities to be as detailed as possible. For instance, in Example 4 we consider words denoting the levels of severity as part of diseases, such as "*nặng*" (severe), "*không hồi phục*" (unrecoverable), "*giai đoạn cuối*" (terminal) and "*mạn* (chronic).