Fine-Grained Named Entities for Corona News

Sefika Efeoglu[1], Adrian Paschke[1,2]

[1]Freie Universitaet Berlin, Takustrasse 9, 14195 Berlin, Germany
[2]Fraunhofer FOKUS, Berlin, Germany

Abstract

Information resources such as newspapers have produced unstructured text data in various languages related to the corona outbreak since December 2019. Analyzing these unstructured texts is time-consuming without representing them in a structured format; therefore, representing them in a structured format is crucial. An information extraction pipeline with essential tasks- named entity tagging and relation extraction- to accomplish this goal might be applied to these texts. This study proposes a data annotation pipeline to generate training data from corona news articles, including generic and domain-specific entities. Named entity recognition models are trained on this annotated corpus and then evaluated on test sentences manually annotated by domain experts evaluating the performance of a trained model. The code base and demonstration are available at https://github.com/sefeoglu/coronanews-ner.git.

Keywords

corona news, named entity recognition, fine-grained entities, contextual embedding

1. Introduction

The coronavirus outbreak has started to spread worldwide from Wuhan, China, the origin of SARS-CoV-2 [1], in late 2019. Local authorities of each country have taken crucial measures -such as tests, vaccines, and mask obligations- in indoor facilities to control the spread of the virus. The authorities give ongoing progress reports on these measures, published on their official web pages and news articles during the pandemic.

After the start of the pandemic, the Covid-19 Open Research Dataset [2] (CORD-19) challenge was declared to convert texts taken from previously published scientific papers in the corona domain into a structured format for downstream applications in March 2020 [1]. Nevertheless, the applications using this corpus fail to identify recent variants of the coronavirus and generic mentions such as organizations and facilities in the news articles. This is because this corpus includes earlier published scientific papers in this domain. Thus, a new up-to-date corpus is needed to analyze all mentions in corona news articles.

This study aims to develop an annotation pipeline that generates annotated training data from newer corona news articles for named entity recognition (NER). After running the annotation

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences

sefika.efeoglu@fu-berlin.de (S. Efeoglu); paschke@inf.fu-berlin.de (A. Paschke)
0000-0002-9232-4840 (S. Efeoglu); 0000-0003-3156-9040 (A. Paschke)

[1]This information was declared by WHO on https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part

[2]https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

pipeline, we leverage the Flair NLP framework [2] and SciBERT [3] to train NER models on the annotated texts and then evaluate the models on test data annotated by domain experts. The corpus in this study is constructed from the corona news articles published in the German news channel "Tagesschau" [3] and are tagged with 23 entity types. The main contribution of this study is to introduce a new corpus from up-to-date corona news articles tagged with gold and silver seeds and a pre-trained NER model (OntoNotes [4]).

The rest of this paper outlines recent works about corona-related text data annotation processes and the NER approaches in the Related Works section. Afterward, the proposed annotation pipeline and the dataset are presented in the Methodology section, and then the experiments carried out are debated in the Evaluation section. Lastly, we summarize our work within this study in the Conclusion section.

2. Related Works

There are several previous attempts to construct a corpus in the corona domain ahead of the Covid-19 pandemic. However, previous corpora in this domain -such as CORD-19 [1] and LitCovid [4]- generated from domain-specific journals need up-to-date information in identifying generic entities and new entities of coronavirus variants.

Wang et al. (2020) introduce a dataset from the CORD-19 corpus using gold seeds created by domain experts, UMLS KB, and NER models (spaCy and SciSpacy) [5]. A NER model using a distant supervision approach is trained on this annotated corpus and evaluated on test sentences annotated by three domain experts. Another study using the CORD-19 corpus is "Automated Text Evidence Mining" supporting data-driven methods for distantly supervised NER and open information extraction [6]. Colic et al.(2020) also propose a pipeline to annotate scientific publications about the Covid-19 pandemic [7]. They leverage the CRAFT corpus annotated with ten entity types from domain-specific ontologies, for example, COVoc4.

Turning to studies using corpora in low-resource languages, Truong et al. (2021) employ a Covid-19 NER model trained on Vietnamese text corpus [8]. The corpus with 35K train and 1K test sentences consists of fully manual annotated texts regarding the Covid-19 situation in Vietnamese. Another study in this field aims to investigate their ability to transfer knowledge between two languages while maintaining necessary features to identify named entities in which datasets are Italian SIRM Covid-19 and English medical records [9].

To sum up, the previous attempts to create an annotated corpus in the corona domain leverage outdated published texts. However, outdated text data is insufficient for recognizing mentions in changing corona news articles.

[3] https://www.tagesschau.de/

[4]The OntoNotes corpus has been constructed from various kinds of data sources to develop information extraction and retrieval application. The details are available on https://catalog.ldc.upenn.edu/LDC2013T19

3. Methodology

3.1. Data and Data Preprocessing

This study collects its data from corona-related news articles published by a German news-channel "Tagesschau" between December 2020 and June 2022, since the outcome of this study will be used in the relation extraction task of the information extraction pipeline later to evaluate the progress of the Covid-19 pandemic in Germany. Due to limited silver seed entities in German, we have translated sentences in these articles into English sentences with Google Translator python library [5]. Before running the annotation pipeline (see Fig. 1) on the data, fundamental text cleaning approaches -e.g, removing unwanted characters like '#' and '*' and unnecessary white spaces- have been applied. After this cleaning process, the pipeline can be applied to unstructured text data to prepare annotated text data for text analytic approaches.

3.2. A Data Annotation Pipeline for Corona News

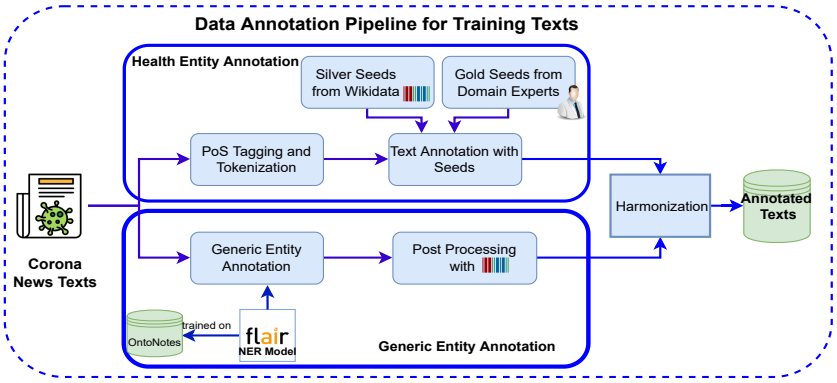

Figure 1: This figure illustrates how a corona news article is annotated with (silver and gold) seed entities and a pre-trained NER model.

This study proposes an annotation pipeline, leveraging silver and gold seeds and a pre-trained NER model (OntoNotes) to prepare annotated texts for text analytics methods like named entity tagging and relation detection between entities in a sentence. This pipeline starts with receiving unstructured text data as input. Then it simultaneously runs two different annotation processes to extract both domain-specific (Health Entity Annotation) and generic (Generic Entity Annotation) entities. Afterwards, it solves conflicts between the annotation results of both processes in the harmonization step by prioritizing the health entity annotation's results where both processes have annotated an entity in the sentence. For instance, if 'Corona' in a sentence is annotated as GPE (Generic) and Coronavirus (Health), the harmonization step will resolve this conflict by assuming 'Corona' as Coronavirus. The pipeline outputs annotated text data for training a NER model.

Health Entity Annotation Process. The corona news articles include domain-specific entities that have not yet been categorized into generic entity types. Therefore, we create

[5] https://pypi.org/project/deep-translator/

gold standard seeds with the assistance of domain experts in domain-specific categories: coronavirus, disease_or_syndrome, sign_or_symptom, and immune_response. These gold seeds can be changed subject to the domain of the corpus before running the annotation pipeline. Furthermore, our domain experts have also provided some generic entities related to corona, for example, vaccine (product), pandemic (event), and family members (group). In addition to these gold seeds, we utilize some silver seeds in these domain-specific categories from Wikidata by running SPARQL and SKOS queries. After obtaining seeds, tokenization and part-of-speech (PoS) tagging are first applied to the corpus. Then, an exact string-matching algorithm is run with the silver and gold seeds for the tokenized sentences in the corpus. Finally, the process identifies all domain-specific named entities defined in both seed sets at the end of this process.

Generic Entity Annotation Process. The corona news articles comprise generic entities like PERSON, FAC, ORG, GPE, and so on. as well, so extracting these entities is also a crucial step for analyzing the news articles. To find the generic entities, we use a NER model pre-trained on OntoNotes having 18 generic entity types [2]. After tagging the generic entities, these entities are sought on Wikidata with the help of SPARQL queries and refine them if found; otherwise, they will remain unmodified. Then, the process outputs annotated texts tagged with generic entities.

4. Evaluation

Experimental Setup. We develop an annotation pipeline for corona news articles and evaluate the performances of NER models trained on these articles. The NER models implemented with the Flair framework [2], utilizing two combinations of embedding types (word and contextual embeddings), and a fine-tuned SciBERT (NER) model [3] use BERT transformer. The hyperparameters of the NER models implemented by Flair are a learning rate of 0.1, 10 epochs, and a batch size of 32, and those of the fine-tuned SciBERT are one epoch and a batch size of 16. The baseline model leverages only Glove word embedding [10]. In contrast, the advanced model has a stack of Glove word embedding and Flair contextual embedding [11] (in forward and backward propagation), providing the model a contextual embedding of a word in a sentence. The numbers of training, validation, and test sentences used in the development of the NER models are 89986, 4999, and 1000, respectively. Before running the pipeline, test sentences were chosen randomly from our initial corpus constructed from the news articles published on Tagesschau. Two domain experts (a medical doctor and a pharmacist) annotated 1000 sentences. Then, to ensure reliability of these test sentences, Fleiss Kappa is calculated as 0.98. These sentences consist of 3126 entities in 23 categories at the end of manual annotation.

Results and Discussion. The results in Table 1 show that the model using the Flair contextual embedding [11] outperforms the other model with only Glove word embedding [10] in most of the newly introduced domain-specific categories of test sentences in terms of the mean micro-F1 score of the five times trained and evaluated models. However, the model leveraging Glove embedding performs better in the 'immune response', and 'disease_or_syndrome' types, since most entities in these categories are single tokens. Moreover, we observe that these types' standard deviation (Std) is pretty high. On the other hand, the fine-tuned SciBERT model's micro F1-score is 0.7765. The entity-specific F1-scores are 0.81 (coronavirus), 0.84 (sign_or_symptom),

0.79 (disease_or_syndrome), 0.8 (immune_response), and 0.85 (group).

Table 1
This table shows the statistical details about mean micro-F1 scores of the NER models, which were trained and evaluated five times. Besides, the table gives the mean micro-F1 scores of new entity types on the models trained with our corona news corpus.

Embedding	Model	Model Std	Coronavirus	Disease or Syndrome	Group	Immune Response	Sign or Symptom
Glove	0.71084	0.003414	0.76522	0.84152	0.80078	0.96364	0.81922
Glove+Flair	0.77162	0.002322	0.78614	0.81214	0.85016	0.83264	0.86562

5. Conclusion

In this study, we propose an annotation pipeline to create annotated texts from the corona news articles for NER. We also contribute with a new up-to-date annotated corpus in the corona domain to identify corona-related mentions on the corona news articles via the NER models. The experiments demonstrate that the models utilizing contextual embedding surpass the model using an only word embedding in terms of micro-F1 score. Besides, the fine-tuned SciBERT model has performed well in the domain-specific entity types. In its next version, we will integrate a spelling-checking API into this pipeline before receiving the texts, since the entities in the articles might have some spelling mistakes after translating them into English.

Acknowledgments

The research presented in this paper was supported in part by the German Federal Ministry of Education and Research (BMBF) project "PANQURA" under grant 03COV03F, in part by the European Union project "FAST-LISA" under grant 101049342.

References

[1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, Cord-19: The covid-19 open research dataset, 2020. URL: https://arxiv.org/abs/2004.10706. doi:10.48550/ARXIV.2004.10706.

[2] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 54–59.

[3] I. Beltagy, K. Lo, A. Cohan, Scibert: Pretrained language model for scientific text, in: EMNLP, 2019. arXiv:arXiv:1903.10676.

[4] Q. Chen, A. Allot, Z. Lu, LitCovid: an open database of COVID-19 literature, Nucleic Acids Research 49 (2020) D1534–D1540. URL: https://doi.org/10.1093/nar/gkaa952. doi:10.1093/nar/gkaa952.

[5] X. Wang, X. Song, B. Li, Y. Guan, J. Han, Comprehensive named entity recognition on cord-19 with distant or weak supervision, 2020. URL: https://arxiv.org/abs/2003.12218. doi:10.48550/ARXIV.2003.12218.

[6] X. Wang, W. Liu, A. Chauhan, Y. Guan, J. Han, Automatic textual evidence mining in covid-19 literature, 2020. URL: https://arxiv.org/abs/2004.12563. doi:10.48550/ARXIV.2004.12563.

[7] N. Colic, L. Furrer, F. Rinaldi, 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, Association for Computational Linguistics, Online, 2020. URL: https://aclanthology.org/2020.nlpcovid19-2.27. doi:10.18653/v1/2020.nlpcovid19-2.27.

[8] T. H. Truong, M. H. Dao, D. Q. Nguyen, 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, Association for Computational Linguistics, Online, 2021, pp. 2146–2153. URL: https://aclanthology.org/2021.naacl-main.173. doi:10.18653/v1/2021.naacl-main.173.

[9] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, M. Esposito, Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set, Applied Soft Computing 97 (2020) 106779. URL: https://www.sciencedirect.com/science/article/pii/S1568494620307171. doi:https://doi.org/10.1016/j.asoc.2020.106779.

[10] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[11] A. Akbik, D. Blythe, R. Vollgraf, Contextual string embeddings for sequence labeling, in: COLING 2018, 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.