Annotating the Pandemic: Named Entity Recognition and Normalisation in COVID-19 Literature

Nico Colic
University of Zurich
colic@ifi.uzh.ch

Lenz Furrer
University of Zurich
furrer@cl.uzh.ch

Fabio Rinaldi
IDSIA
fabio@idsia.ch

Abstract

The COVID-19 pandemic has been accompanied by such an explosive increase in media coverage and scientific publications that researchers find it difficult to keep up.

We are presenting a publicly available pipeline to perform named entity recognition and normalisation in parallel to help find relevant publications and to aid in downstream NLP tasks such as text summarisation. In our approach, we are using a dictionary-based system for its high recall in conjunction with two models based on BioBERT for their accuracy. Their outputs are combined according to different strategies depending on the entity type. In addition, we are using a manually crafted dictionary to increase performance for new concepts related to COVID-19.

We have previously evaluated our work on the CRAFT corpus, and make the output of our pipeline available on two visualisation platforms.

1 Introduction

The body of scientific literature is growing at an unprecedented rate, and this is particularly evident in the response of the biomedical research community to the 2020 COVID-19 pandemic. Several platforms have been established to track publications related to COVID-19, most prominently the COVID-19 Open Research Dataset (CORD-19)[1], a collaboration of the US Government and multiple other organisations, the LitCovid dataset, maintained by the NIH, which indexes papers published on PubMed related to the pandemic (Chen et al., 2020), or Novel Coronavirus Research Compendium (NCRC)[2], which contains 800 publications selected manually for their originality and quality.

[1] semanticscholar.org/cord19
[2] ncrc.jhsph.edu/

In this publication, we are processing the articles of the LitCovid dataset, which at the time of writing contains over 20 000 publications related the 2020 COVID-19 pandemic only, showing growth at a steady rate since its beginning.

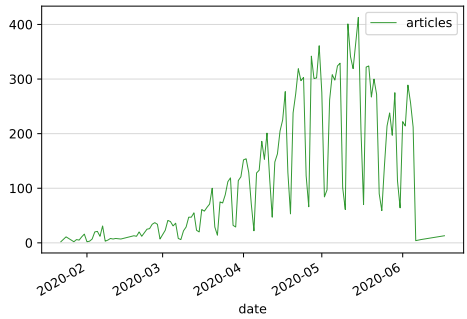

Figure 1: Publications per day included in LitCovid

The flurry of news and public discussions about the pandemic, which includes a substantial amount of fake news, has been termed *"infodemic"*. However, the term could be applied also to the rapid growth of reports and publications pertaining the disease (see Figure 1). Interestingly, this growth pattern seems to resemble that of the spread of the disease in western countries (with a delay of one to two months).

While the growth is not exponential as it has occasionally been reported, it is still far beyond what virologists and medical scientists can manually process. This is an exacerbation of a general problem in biomedical research, where researchers cannot keep up with the growth of literature that pertains to their research, and need to resort to named entity recognition (NER), named entity normalisation (NEN) and text summarisation technologies to identify relevant publications (Lu, 2011).

In NER, entities of interest are identified as text spans in free text; and then, in NEN, mapped to

unique IDs in a controlled vocabulary. They constitute a fundamental step for other down-stream text processing tasks, on one hand; but are also a means to its own end, allowing publications to be indexed by the entities they contain, on the other hand.

In previous research, we have shown that we can obtain better results by performing NER and NEN in parallel rather than sequentially, avoiding propagation of errors between the steps. We are building on this previous research and add a further processing step to find terms specific to COVID-19.

2 Related Work

In March 2020, the US White House collaborated with the National Library of Medicine, the Allen Institue for Artificial Intelligence and other private companies to create the CORD-19 corpus (Wang et al., 2020a), and with it a set of 18 challenges such as *What do we know about COVID-19 risk factors?* for data scientists to participate in, hosted on Kaggle[3].

The response of the text mining community to the pandemic and such shared tasks has been enormous, producing a wide array of webservices, machine learning models and databases; usually adapting existing frameworks to suit the pandemic. Wang et al. (2020c), for example, are retraining SciSpacy on the CORD-19 corpus to improve its NER performance.

Some research has already been directed at downstream tasks, using a simple dictionary-based NER method as a base to perform entity relation extraction (Rao et al., 2020; Wang et al., 2020b), to create a knowledge base (Khan et al., 2020) or for summarisation systems (Gutierrez et al., 2020; Kieuvongngam et al., 2020).

The problem of NER and NEN in the biomedical domain, generally, has traditionally been approached with pipelines, using rules or dictionaries (Campos et al., 2013; D'Souza and Ng, 2015). More recently, however, machine learning using various architectures such as LSTMs or CRFs have become more popular (Leaman et al., 2013; Habibi et al., 2017).

In this vein, it has been suggested to approach NER and NEN simultaneously (ter Horst et al., 2017; Lou et al., 2017), which is similar to the approach that we follow.

The authors of the LitCovid data set, which we process in the present work, also perform NER and NEN on the dataset using PubTator (Wei et al., 2019). In their work, they annotate for 6 entity types (genes, diseases, chemicals, mutations, species and cells) and use a different architecture for every single type. For example, they use a linear classifier for annotating diseases (Leaman and Lu, 2016), and a BERT-based transformer for finding chemicals. This differs fundamentally from our approach, where we employ the same architecture for all entity types. Furthermore, apart from the NCBI Taxonomy, we are using different controlled vocabularies for entity normalisation for all types.

3 Pipeline

In our approach, we build on our previous efforts where we use a parallel architecture to perform NER and NEN simultaneously (Furrer et al., 2019a, 2020). Traditionally, NER and NEN are performed after each other, which means that spans of mentions of entities are identified first, and then mapped to the corresponding entry in a controlled vocabulary. This approach has the drawback that errors made in the first step are irrecoverably propagated to the second stage.

In our approach, however, we perform those two steps simultaneously, and were able to show that it outperforms the traditional approach (Furrer et al., 2019a). We are using BioBERT, a pre-trained language model, which we trained on the CRAFT corpus, a collection of nearly 100 full-text medical articles manually annotated for 10 different medical entity types. We have evaluated our approach using the CRAFT corpus, and obtained F1-scores between 0.74 and 0.92 depending on the entity type.

To improve our results on COVID-19 literature, we are adding an additional step of post-annotating our results using a manually crafted dictionary specific to COVID-19.

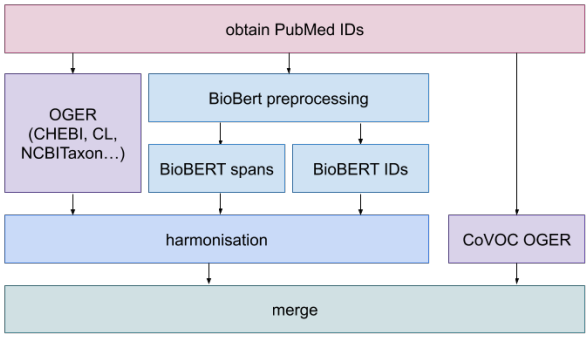

Figure 2: Overall structure of the pipeline

[3] https://bit.ly/384VgBQ

3.1 Vocabularies

The dataset is annotated for entities coming from 10 different ontologies as they are used in the CRAFT corpus, such as *Chemical Entities of Biological Interest* (**CHEBI**) or the *NCBI Taxonomy* (**NCBITaxon**). Additionaly, we employ the manually curated, COVID-19 specific terminology COVoc[4], containing over 250 terms. We are using these ontologies because we were able to test our performance using the CRAFT corpus, and because they provide extensive coverage over the biomedical domain (Cohen et al., 2017).

3.2 OGER

OGER is a dictionary-based look-up tool using an efficient fuzzy matching algorithm (Furrer et al., 2019b). Relying on a dictionary mapping relevant entities to their ID, its performance depends on the dictionary's quality and extent, which manually or automatically curated ontologies such as CHEBI provide. It thus requires no training, and can detect entities that an example-based system would miss if they are not present in the training data, provided they are present in the dictionary.

3.3 BioBERT

BERT is a multi-layer transformer trained on the English Wikipedia and BookCorpus (Devlin et al., 2018). While it is trained to predict whether a sentence follows another and randomly blacked out words, the resulting language model can be fine-tuned for different tasks, such as NER (Hakala and Pyysalo, 2019) and NEN, or adapted for different domains through further training. BioBERT is the result of training BERT on PubMed articles, making it useful for biomedical applications (Lee et al., 2020; Sun and Yang, 2019).

We have used BioBERT and trained it further on the CRAFT corpus to build a span prediction and an ID prediction model. The span predictor produces IOBES labels, and is used in conjunction with OGER to provide ID labels. The ID predictor also conceptualises NEN as a sequence tagging problem and works like a classical NER model, but with the output tagset extended to cover all possible concept labels.

The ID predictor thus predicts spans and IDs directly, making the use of other models theoretically superfluous. However, it suffers from the fact that it cannot predict concepts not seen during training and that it does not perform well for tokens that occur both in general domain language and in biomedical entities (such as *I* in *hexokinase I*). By using the span prediction model in conjunction with OGER, too, we alleviate these shortcomings.

3.4 Harmonising, annotating for COVID-19, merging

For conflicting or overlapping annotations between the BioBERT span and ID classifiers as well as OGER, we were able to show in our previous work that the optimal merging strategy depends on the entity type in question (Furrer et al., 2020). In this step, we take these findings into account when deciding which system's output to prioritise for the final output. If a span prediction is given preference, the ID label as produced by OGER as described in Section 3.3 is used.

In a last step, we run OGER again to produce an additional layer of annotations for terms specific to COVID-19 using the COVoc vocabulary. In this way we hope to be able to maintain the accuracy of our models for the established vocabularies, while allowing for rapid changes to be made to the set of entities specific to the pandemic without having to retrain the BioBERT modules.

The outputs are then merged for all entity types, and converted to various formats.

4 Results

So far, with our pipeline we have processed over 25 000 abstracts from PubMed and 7883 full-text articles from PMC, with a total amount of over 400 000 and 900 000 annotations, respectively (see Table 1).

With our pipeline, we are able to continuously process new articles that are added to the LitCovid dataset, and distribute our annotations in the following ways:

- PubAnnotation and EuroPMC
- Our own webservice using BRAT
- Freely downloadable files

The OGER annotations can be obtained through an API[5]. The code to run the pipeline[6], its outputs[7] as well as the CRAFT-trained BioBERT models[8]

[4] bit.ly/2BK0u9W

[5] https://bit.ly/2Vrbekw

[6] github.com/Aequivinius/covid

[7] https://bit.ly/3eMylOq

[8] doi.org/10.5281/zenodo.3822363

vocabulary	PM abstracts	PMC articles
CoVoc	165668	261287
UBERON	79899	204355
NCBITaxon	67278	147524
GO_BP	34510	84604
CHEBI	30720	99673
PR	12319	48471
GO_CC	7656	28738
CL	7332	28849
SO	6801	25017
MOP	449	2559
GO_MF	73	260
total	**412 705**	**931 337**

Table 1: Annotations per vocabulary for PubMed and PMC

are publicly available, and with some effort could be modified using OGER's format conversion to process other dataset such as CORD-19.

4.1 Online Repositories

PubAnnotation is an online repository for annotations on PubMed articles, (Kim et al., 2015, 2019), which also features the annotation visualisation engine **TextAE** (see Figure 3). Europe PMC is a repository of publications akin to PubMed, but also allows display of annotations (Consortium, 2015). We uploaded our annotations to both services.

4.2 BRAT

On our own infrastructure[9], we host an instance of BRAT, which visualises annotations in a similar fashion as PubAnnotation (Stenetorp et al., 2012).

4.3 Downloads

To further facilitate down-stream tasks, we provide our annotations in the most frequently used annotation formats[10]: `.txt`, CoNLL `.tsv` and BioC `.json`.

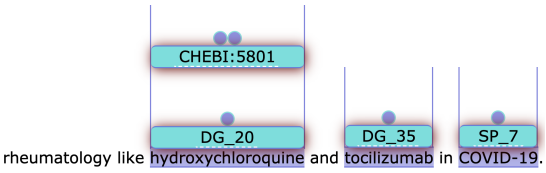

Figure 3: Annotations visualised by PubAnnotation's TextAE.

[9] bit.ly/3eITn0o
[10] bit.ly/386BbuN

5 Discussion

Tools that automatically process literature related COVID-19 generally fall into two broad categories: Systems that follow some sort of text summarisation approach, and NER+NEN systems.

Much attention has been directed at previously mentioned Kaggle challenge, for which over 1500 solutions have been submitted, ranging from statistical data exploration to a full clustering of the literature. One of the top submissions[11], for example, attempts to identify risk factors of COVID-19 by applying unsupervised topic modeling algorithms. Such approaches are very common among the submissions, but suffer from a high number of false positives.

Similarly, platforms that allow browsing corpora of COVID-19 papers such as COVIDScholar[12] and the BERT-driven COVID-19 Research Explorer[13] rely on word embeddings and other unsupervised algorithms to find matching publications or even passages in publications. For the latter, the authors attempt to go beyond traditional document retrieval, and employ an automatically generated corpus to fuel their question answering learning (Ma et al., 2020). However, such approaches lack the precision typical NER+NEN-driven approaches offer, and don't perform particularly well at matching entity synonyms due to their representation as high-recall word vectors rather than precisely matched entities.

For example, both applications yield different results for either *Angiotensin converting enzyme 2* or *ACE2*, even though the terms are equivalent (and link to the same entry in the Protein Ontology). Repositories that perform controlled vocabulary NEN such as KnetMiner, for example, avoid this error (Hassani-Pak et al., 2020).

Services exploring the scientific literature still fall in *either* of the two camps, and thus fail to exploit the high precision benefits NER+NEN offers and the variety of applications text summarisation approaches afford simultaneously.

Given the urgency of the pandemic, there is currently a lack of resources that allow evaluation of work on the COVID-19 literature, and we hope to be able to test the efficacy of our own work in the future when such resources become available.

[11] bit.ly/2VkN6QP
[12] covidscholar.org/
[13] bit.ly/3fWNOLG

References

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. A modular framework for biomedical concept recognition. *BMC bioinformatics*, 14(1):1–21.

Q. Chen, A. Allot, and Z. Lu. 2020. Keep up with the latest coronavirus research. *Nature*, 579(7798):193.

K Bretonnel Cohen, Karin Verspoor, Karën Fort, Christopher Funk, Michael Bada, Martha Palmer, and Lawrence E Hunter. 2017. The Colorado Richly Annotated Full Text (CRAFT) corpus: Multi-model annotation in the biomedical domain. In *Handbook of Linguistic Annotation*, pages 1379–1394. Springer.

Europe PMC Consortium. 2015. Europe pmc: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, 43(D1):D1042–D1048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302.

Lenz Furrer, Joseph Cornelius, and Fabio Rinaldi. 2019a. UZH@CRAFT-ST: a sequence-labeling approach to concept recognition. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 185–195.

Lenz Furrer, Joseph Cornelius, and Fabio Rinaldi. 2020. Parallel sequence tagging for concept recognition. *arXiv preprint arXiv:2003.07424*.

Lenz Furrer, Anna Jancso, Nicola Colic, and Fabio Rinaldi. 2019b. OGER++: hybrid multi-type entity recognition. *Journal of Cheminformatics*, 11(1):7.

Bernal Jimenez Gutierrez, Juncheng Zeng, Dongdong Zhang, Ping Zhang, and Yu Su. 2020. Document classification for covid-19 literature.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China. Association for Computational Linguistics.

Keywan Hassani-Pak, Ajit Singh, Marco Brandizi, Joseph Hearnshaw, Sandeep Amberkar, Andrew L Phillips, John H Doonan, and Chris Rawlings. 2020. KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *bioRxiv*.

Hendrik ter Horst, Matthias Hartung, and Philipp Cimiano. 2017. Joint entity recognition and linking in technical domains using undirected probabilistic graphical models. In *International Conference on Language, Data and Knowledge*, pages 166–180. Springer.

Junaed Younus Khan, Md Khondaker, Tawkat Islam, Iram Tazim Hoque, Hamada Al-Absi, Mohammad Saifur Rahman, Tanvir Alam, and M Sohel Rahman. 2020. Covid-19base: A knowledgebase to explore biomedical entities related to covid-19. *arXiv preprint arXiv:2005.05954*.

Virapat Kieuvongngam, Bowen Tan, and Yiming Niu. 2020. Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *arXiv preprint arXiv:2006.01997*.

Jin-Dong Kim, Kevin Bretonnel Cohen, and Jung-jae Kim. 2015. Pubannotation-query: a search tool for corpora with multi-layers of annotation. In *BMC Proceedings*, volume 9, pages 1–3. BioMed Central.

Jin-Dong Kim, Yue Wang, Toyofumi Fujiwara, Shujiro Okuda, Tiffany J Callahan, and K Bretonnel Cohen. 2019. Open agile text mining for bioinformatics: the PubAnnotation ecosystem. *Bioinformatics*, 35(21):4372–4380.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov models. *Bioinformatics*, 32(18):2839–2846.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics*, 33(15):2363–2371.

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2020. Zero-shot neural retrieval via domain-targeted synthetic query generation. *arXiv preprint arXiv:2004.14503*.

Aditya Rao, VG Saipradeep, Thomas Joseph, Sujatha Kotte, Naveen Sivadasan, and Rajgopal Srinivasan. 2020. Text and network-mining for covid-19 intervention studies.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Cong Sun and Zhihao Yang. 2019. Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020a. CORD-19: the Covid-19 open research dataset. *ArXiv*.

Xuan Wang, Weili Liu, Aabhas Chauhan, Yingjun Guan, and Jiawei Han. 2020b. Automatic textual evidence mining in covid-19 literature. *arXiv preprint arXiv:2004.12563*.

Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020c. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *arXiv preprint arXiv:2003.12218*.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593.