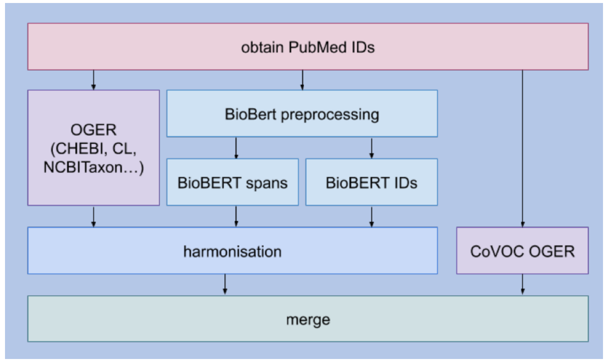
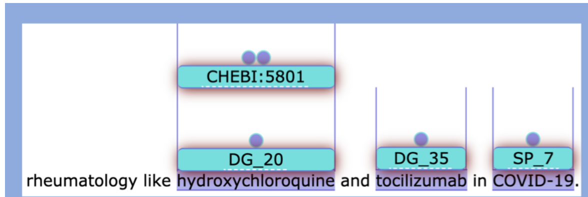
Annotating the Pandemic: NER and NEN of LitCovid

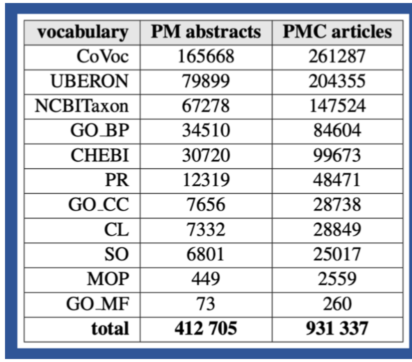
Entity recognition and normalization on COVID-19 literature using a CRAFT-trained BioBERT model for its precision and our dictionary-based tool for its recall.





rheumatology like hydroxychloroquine and tocilizumab in COVID-19.

Pipeline output is uploaded to PubAnnotation, for example, where it is visualized via TextAE. We're also uploading our results to EuroPMC, our own webserver using BRAT, and allowing downloads in JSON and CoNLL TSV for downstream tasks.



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**

Annotations per entity type for PubMed (abstracts) and PubMed Central (full articles)

LitCovid is a dataset of 20 000 PubMed articles related to COVID-19. We are using our pipeline, which performed with F1-score of 0.74 and 0.92 on the CRAFT corpus, depending on entity type (chemical, disease...). Output of models (BioBERT and OGER, our dictionary-based tool) was merged according to different strategies determined most effective in previous work depending on the entity type. The BioBert models produce either ID or span annotations. In the latter case, the ID of the entity was supplied by OGER. This approach helps to optimize both recall and precision. Then, another run of OGER with a hand-crafted dictionary for terms specific to COVID-19, allowing us to make quick changes without retraining models.

Universität Zürich UZH

BioMeXT
www.biomext.org

paper on OpenReview: