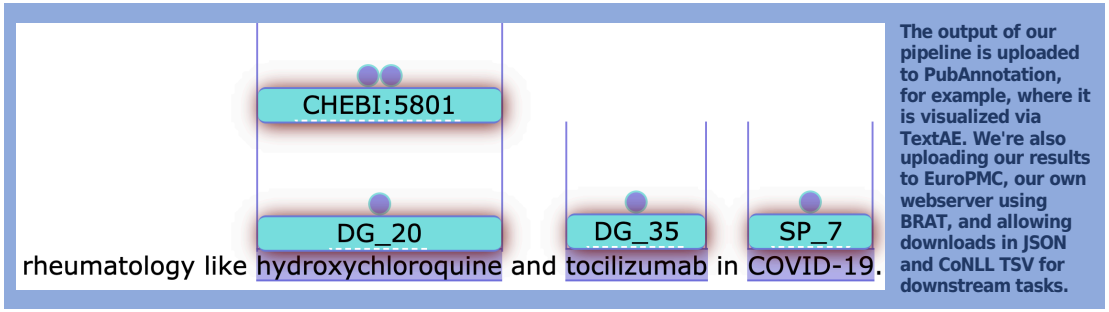
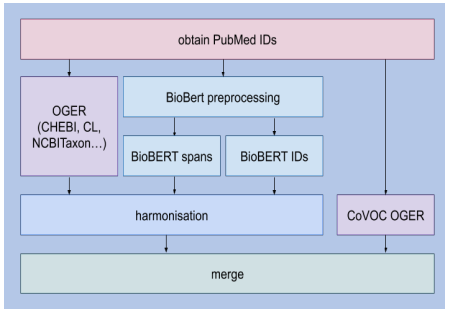
Annotating the Pandemic: NER and NEN of LitCovid

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

All the data can be found online at **covid19.nlp.idsia.ch/lbd.html**



The output of our pipeline 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.



LitCovid is a dataset of PubMed articles related to COVID-19. We use our pipeline, which performs with a F1-score of 0.74 - 0.92 on CRAFT, depending on entity type (chemical, disease…). Our pipeline merges the output of different models (BioBERT and OGER, our dictionary-based tool) according to the strategy determined most effective in previous work for the respective 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 the models.

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 (20k abstracts) and PMC (5k full articles)

Universität Zürich[UZH]

USI/SUPSI Istituto Dalle Molle di studi sull'intelligenza artificiale IDSIA

BioMeXT
www.biomext.org

SIB Swiss Institute of Bioinformatics

Furrer@cl.uzh.ch
Fabio.Rinaldi@idsia.ch
Colic@ifi.uzh.ch