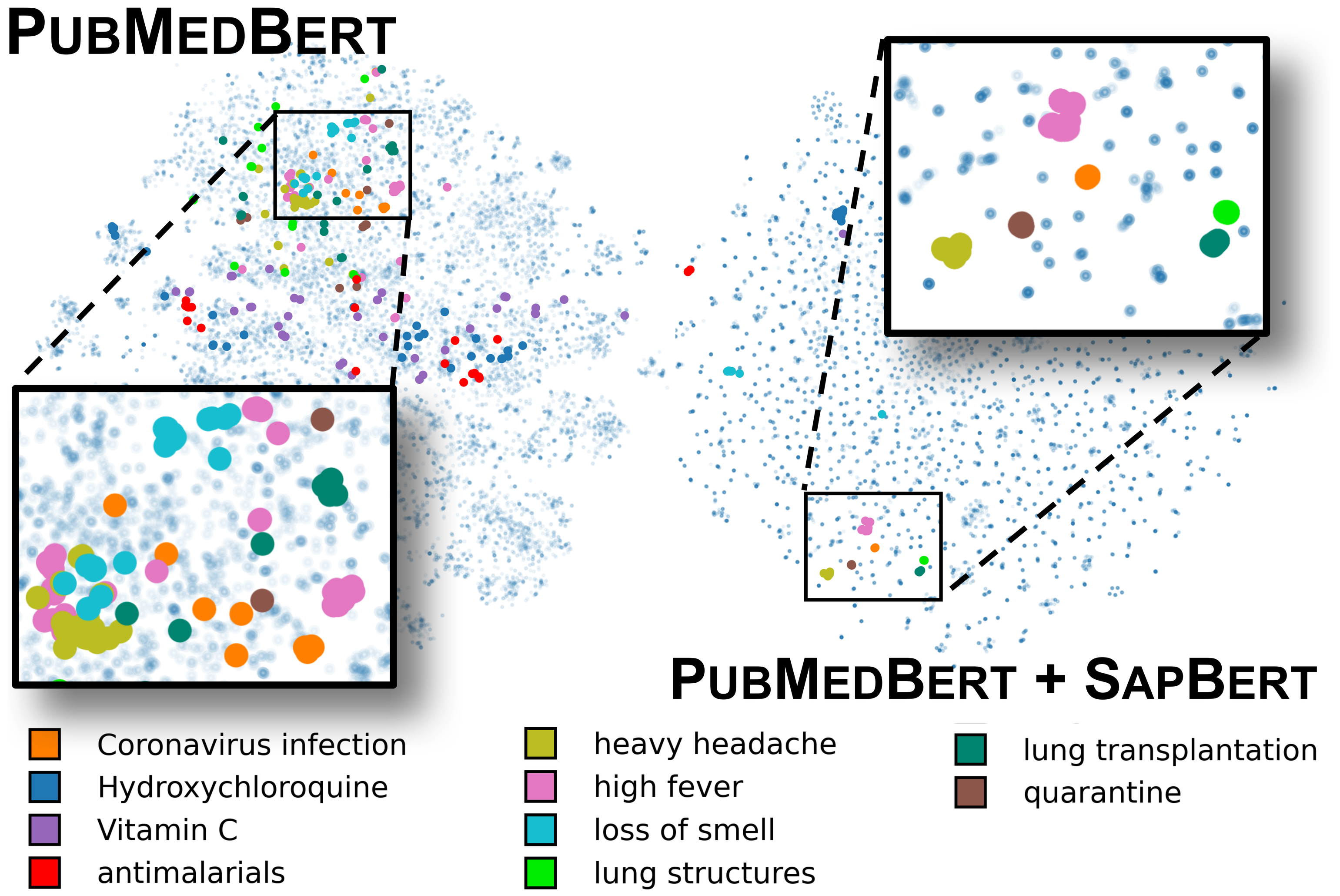
Self-alignment Pretraining for Biomedical Entity Representations

Fangyu Liu [1], Ehsan Shareghi [1,2], Zaiqiao Meng [1], Marco Basaldella[1], Nigel Collier[1]

[1]University of Cambridge, UK [2]University College London, UK



0 Study object: biomedical entities

What is a biomedical entity?

- a single word (e.g. *fever*)
- a compound (e.g. *SARS-COV-2*)
- a phrase (e.g. *abnormal retinal vascular development*)

1 Challenge: heterogeneous naming

Biomedical names referring to the same concept have drastically different surface forms:

- *Hydroxychloroquine*
- *Oxichlorochine* (alternative spelling)
- *HCQ* (social media)
- *Plaquenil* (drug name)
- ......

This is a major challenge for MLM-style pretraining. How do we cope this?

2 Pretraining resource: UMLS (a gigantic KG)

UMLS is the largest interlingua of biomedical ontologies, containing a comprehensive collection of biomedical synonyms in various forms. Some stats: **4M+ concepts** and **10M+ synonyms**, stemming from **over 150 controlled vocabularies**. We design a metric learning framework that self-aligns synonym representations belonging to the same UMLS concept.

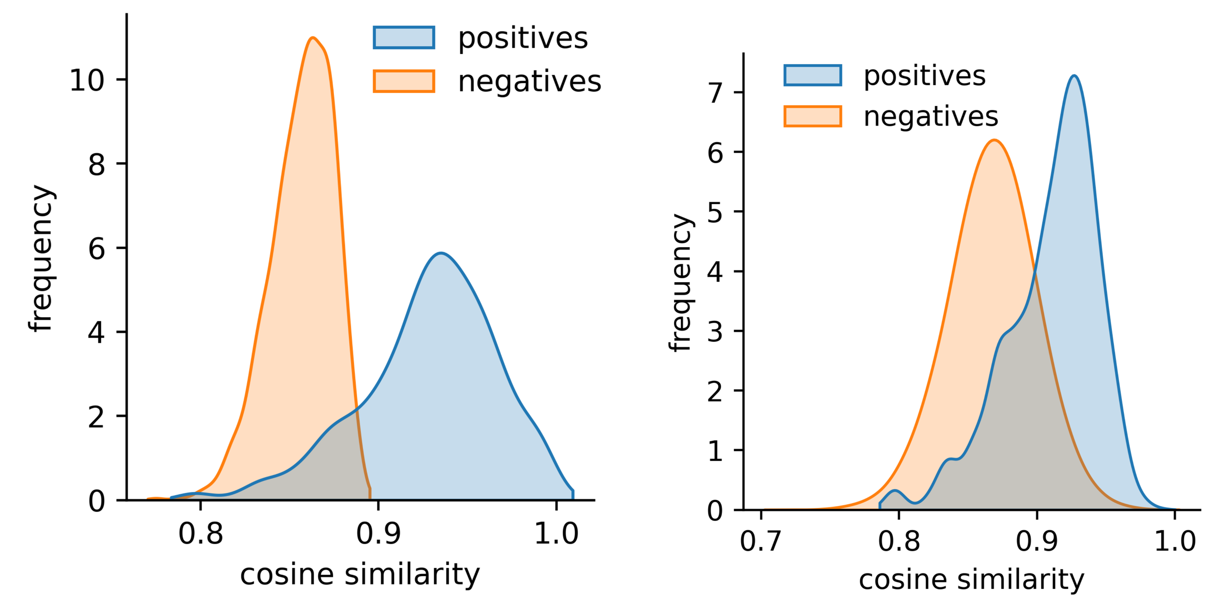
3 Method: self-alignment pretraining

The goal of the self-alignment is to learn a function $f(\cdot;\theta):\mathcal{X}\rightarrow\mathbb{R}^d$ s.t. the similarity $\langle f(x_i), f(x_j)\rangle$ is high if x_i, x_j are synonyms and low otherwise. A sampling procedure selects the informative pairs of training samples and uses them in the pairwise metric learning loss function (introduced below).

Online hard pairs mining:

$$\|f(x_a)-f(x_p)\|_2 < \|f(x_a)-f(x_n)\|_2+\lambda. \quad (1)$$

Intuition: most of *Hydroxychloroquine*'s variants are easy: *Hydroxychlorochin*, *Hydroxychloroquine (substance)*, *Hidroxicloroquina* and etc., but a few can be very hard: *Plaquenil* and *HCQ*. This step forces the model to focus only on the informative examples. Shown below: cosine similarity of pos./neg. pairs before (left) and after (right) applying online hard mining.

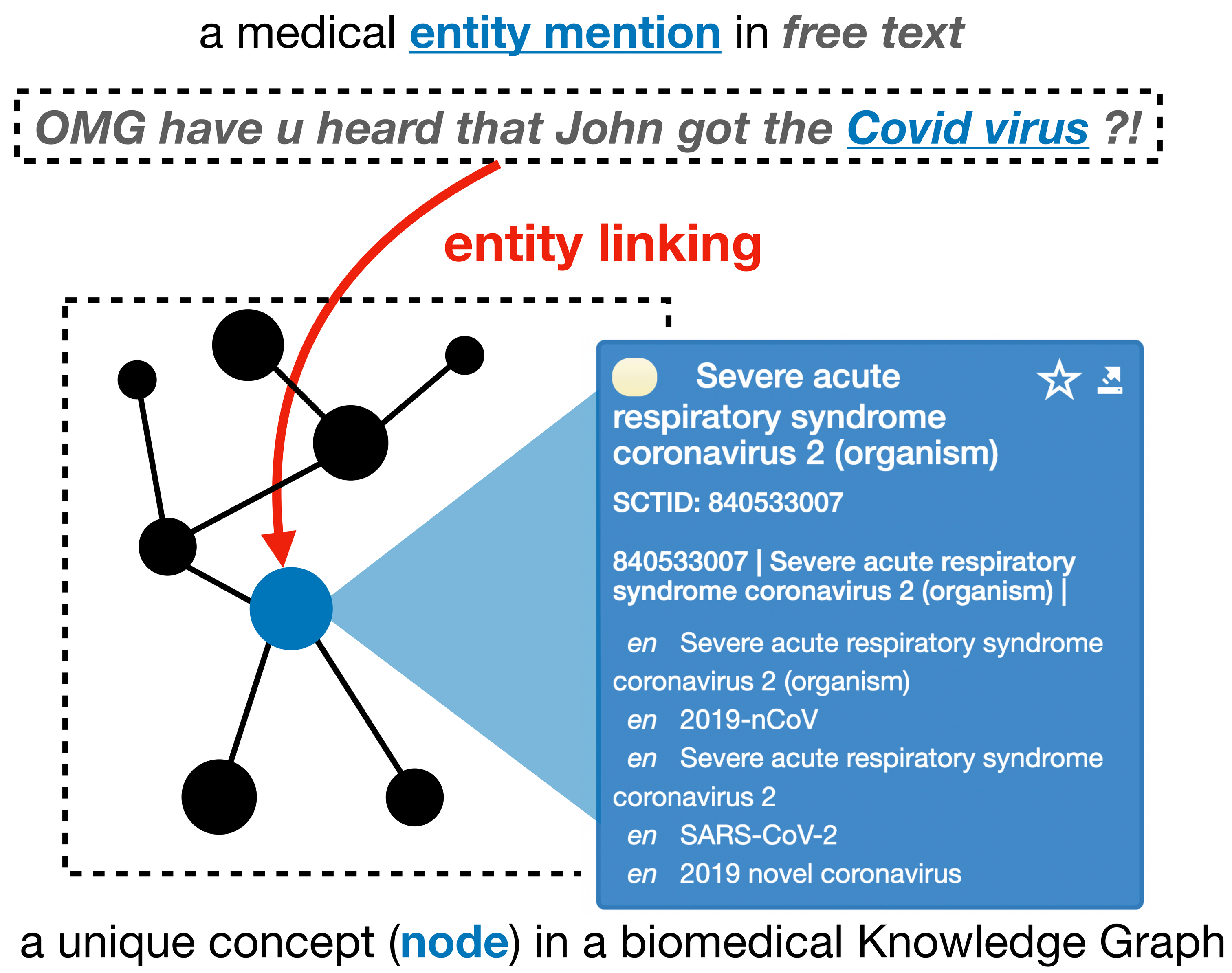


Multi-Similarity loss (MS loss):

$$\mathcal{L}=\frac{1}{|\mathcal{X}_b|}\sum_{i=1}^{|\mathcal{X}_b|}\left(\frac{1}{\alpha}\log\left(1+\sum_{n\in\mathcal{N}_i}e^{\alpha(\mathbf{S}_{in}-\epsilon)}\right)+\frac{1}{\beta}\log\left(1+\sum_{p\in\mathcal{P}_i}e^{-\beta(\mathbf{S}_{ip}-\epsilon)}\right)\right). \quad (2)$$

Intuition: see paper for details.

4 Evaluation: entity linking



Quantitative results (accuracy across 6 data sets):

domain→	scientific				social media	
model↓, data set→	D1	D2	D3	D4	D5	D6
vanilla BERT	67.6	81.4	79.8	39.6	38.2	40.4
+ SapBERT	91.6	92.7	96.1	52.5	68.4	59.5
BioBERT	71.3	79.8	74.0	24.2	41.4	35.9
+ SapBERT	91.0	93.3	95.5	97.6	72.4	63.3
PubMedBERT	77.8	89.0	93.0	43.9	42.5	46.8
+ SapBERT	92.0	93.5	96.5	50.8	70.5	65.9
supervised SOTA	91.1	93.2	<u>96.6</u>	OOM	87.5	<u>79.0</u>
PubMedBERT	77.8	89.0	93.0	43.9	42.5	46.8
+ SapBERT	92.0	<u>93.5</u>	96.5	**50.8**	70.5	65.9
+ SapBERT (fine-tuned)	<u>92.3</u>	93.2	96.5	<u>50.4</u>	**89.0**	**81.1**
BioSyn	91.1	93.2	<u>96.6</u>	OOM	82.6	71.3
+ (init. w/) SapBERT	**92.5**	**93.6**	**96.8**	OOM	<u>87.6</u>	77.0

Table 1: The gradient of green indicates the improvement comparing to the base model (the deeper the more). Blue and red denote unsupervised and supervised models. **Bold** and <u>underline</u> denote the best and second best results in the column.