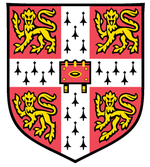
SapBERT: Self-Alignment Pretraining for Biomedical Entity Representations

(Accepted to NAACL 2021.)

Language Technology Lab, University of Cambridge, UK

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, Nigel Collier



UNIVERSITY OF CAMBRIDGE

April 16th, 2021, AstraZeneca (online)

Intro

1 Study Object: Biomedical Entities

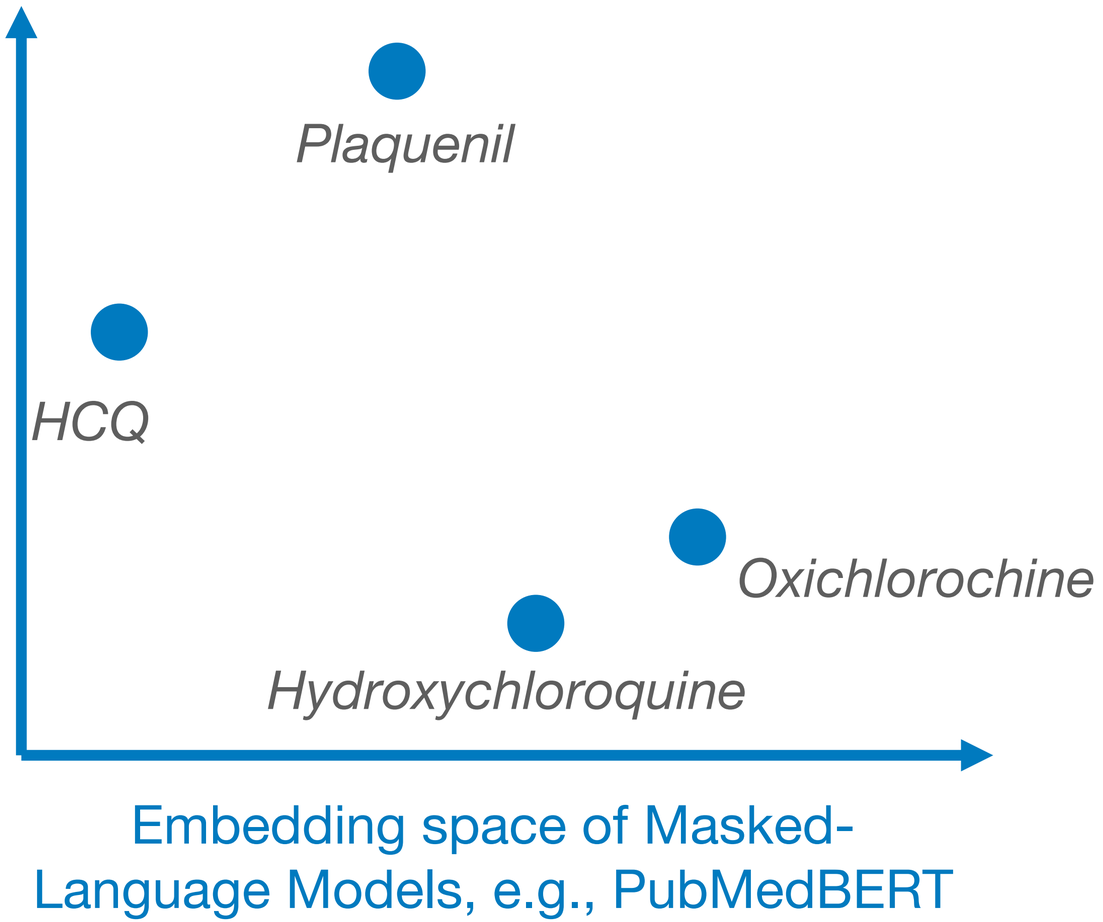
What counts as a biomedical entity?

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

2 Challenge: Heterogeneous Naming

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

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

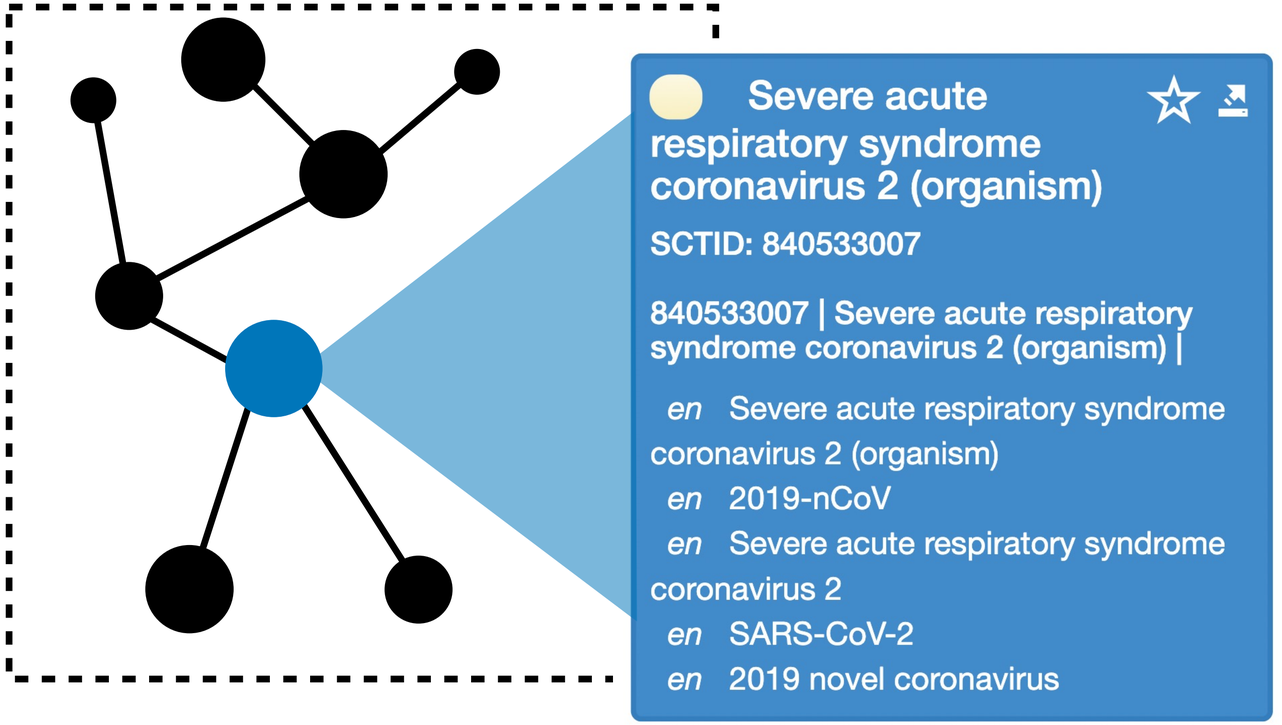


Embedding space of Masked-Language Models, e.g., PubMedBERT

3 Solution: Knowledge Injection from Ontologies

We use UMLS, the largest interlingua of biomedical ontologies:

- 4M+ concepts
- 10M+ synonyms
- 150+ controlled vocabularies (e.g. SNOMED, RxNORM, …)

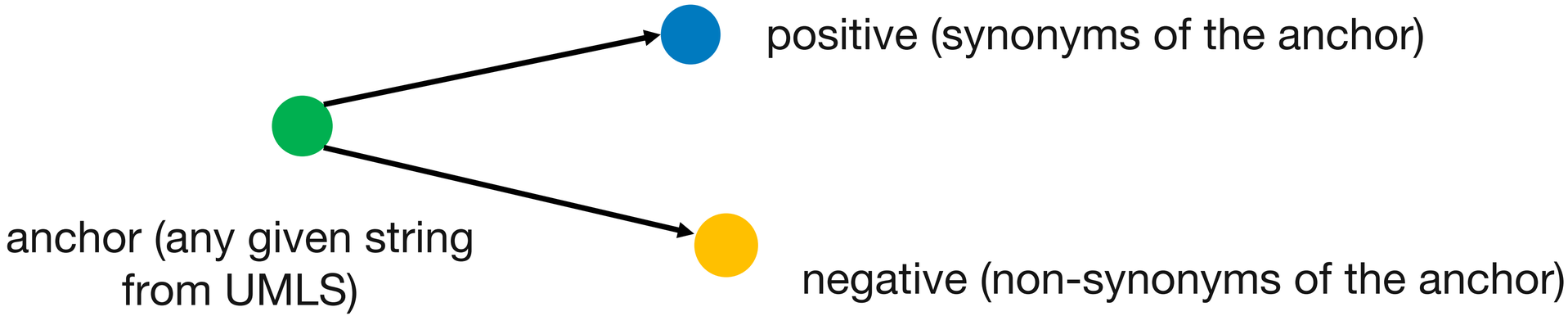


Synonym relations extracted from UMLS can inform language models' representations

Method

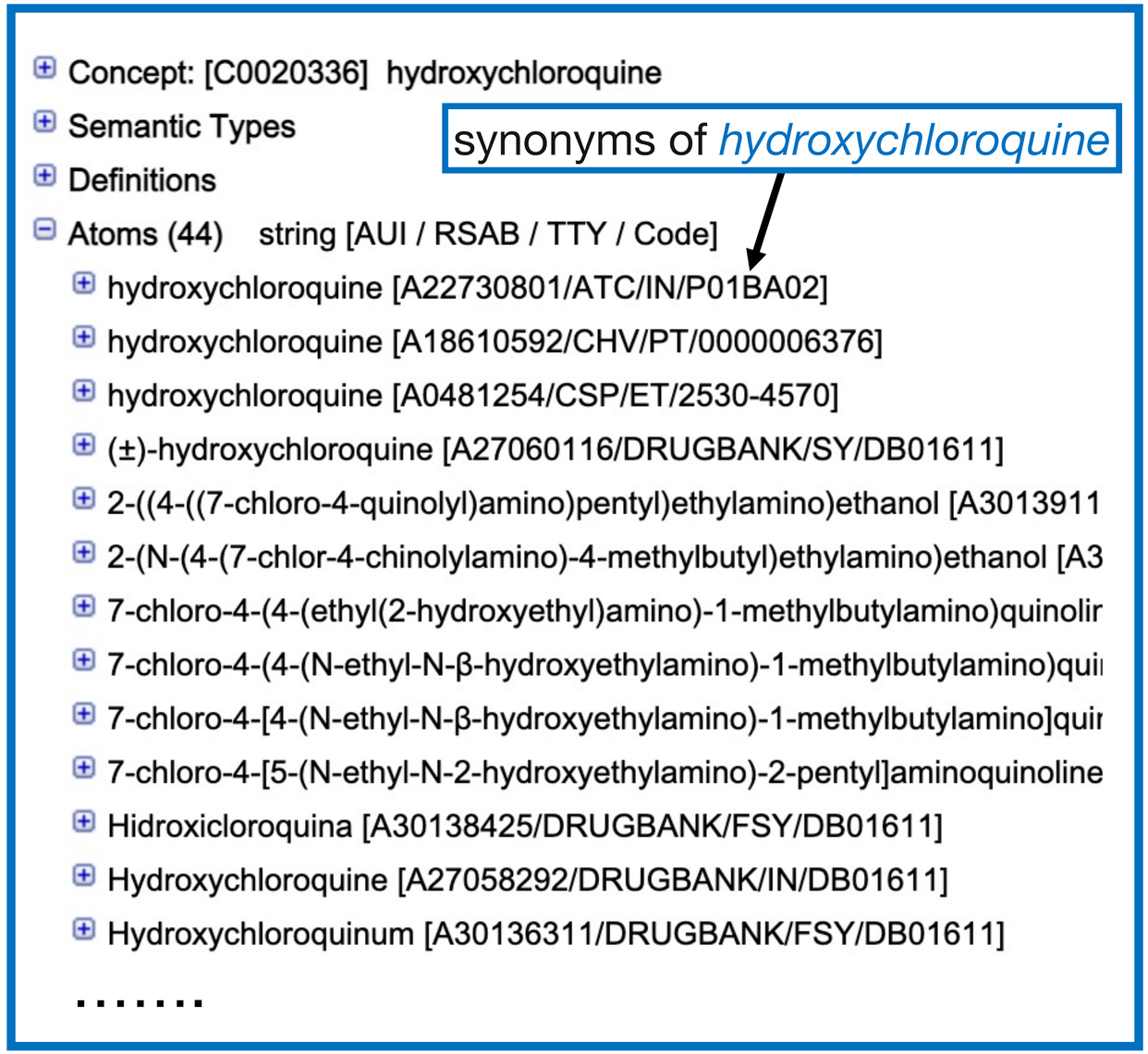
4 Method: Self-Alignment Pretraining

Constructing positives and negatives



4 Method: Self-Alignment Pretraining

Technical challenge: UMLS is gigantic, but not always informative



Most of *hydroxychloroquine*'s variants are easy:

- *Hydroxychlorochin*
- *Hydroxychloroquine (substance)*
- Hidroxicloroquina
- ……

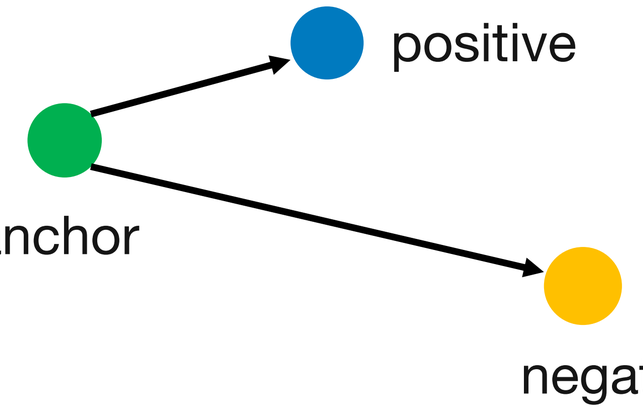
But a few can be very hard:

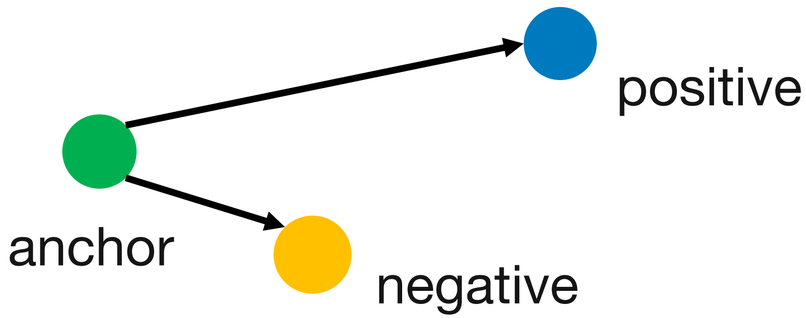
- *HCQ*
- *Plaquenil*
- ……

Can we focus on/learn more from the hard/informative examples?

4 Method: Self-Alignment Pretraining

Techniques: (1) smart online sampling

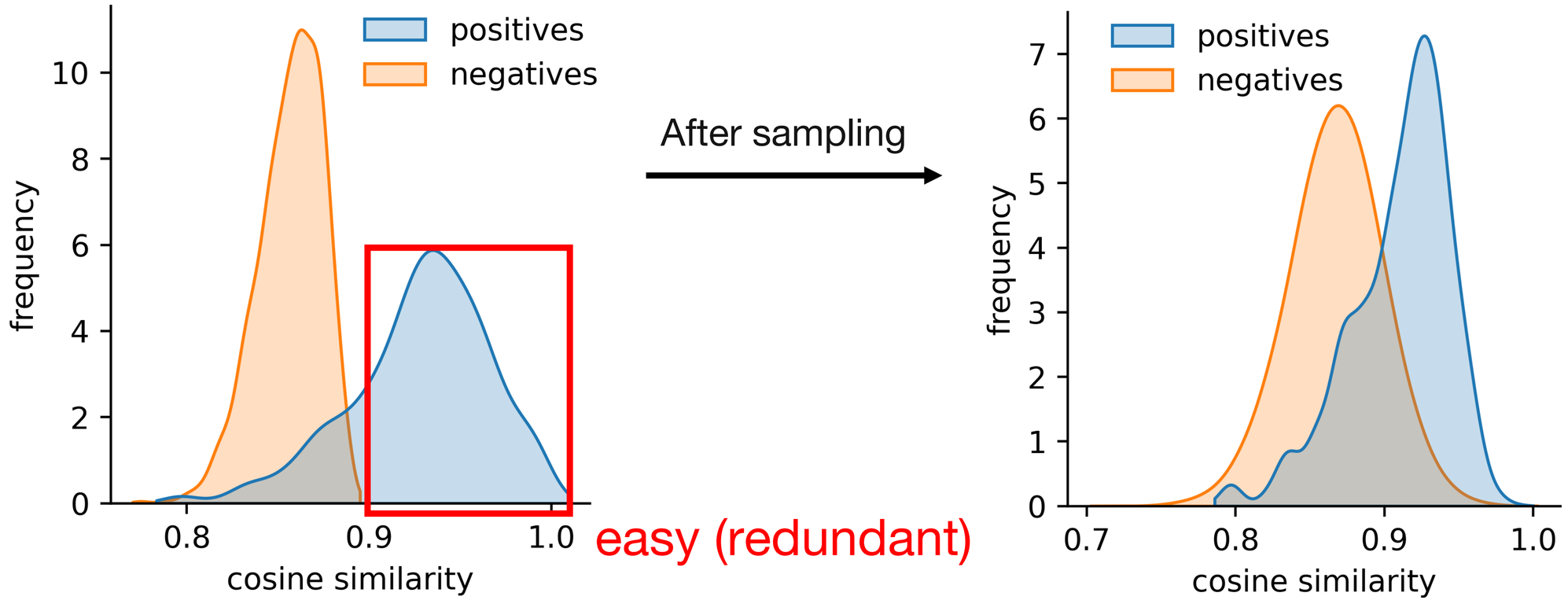




$$\|f(x_a) - f(x_p)\|_2 > \|f(x_a) - f(x_n)\|_2 + \lambda$$

4 Method: Self-Alignment Pretraining

Techniques: (1) smart online sampling



4 Method: Self-Alignment Pretraining

Techniques: (2) multi-similarity loss

push away negatives

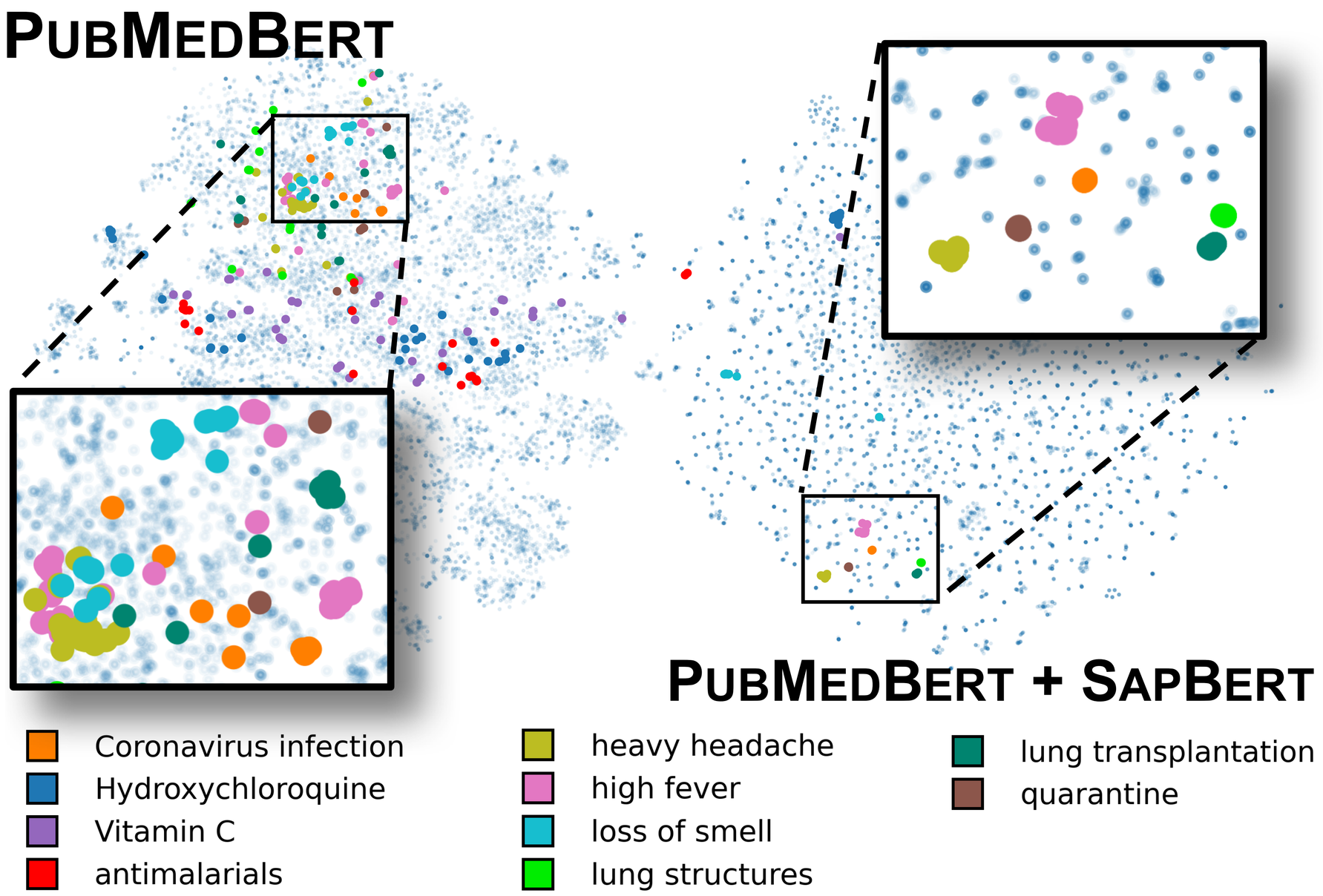
$$\mathcal{L} = \frac{1}{|\mathcal{X}_b|} \sum_{i=1}^{|\mathcal{X}_b|} \left(\frac{1}{\alpha} \log \left(1 + \sum_{n \in \mathcal{N}_i} e^{\alpha(\mathbf{S}_{in} - \epsilon)} \right) + \frac{1}{\beta} \log \left(1 + \sum_{p \in \mathcal{P}_i} e^{-\beta(\mathbf{S}_{ip} - \epsilon)} \right) \right),$$

cluster positives

5 Evaluations

5 Evaluations

(1) T-SNE visualisations



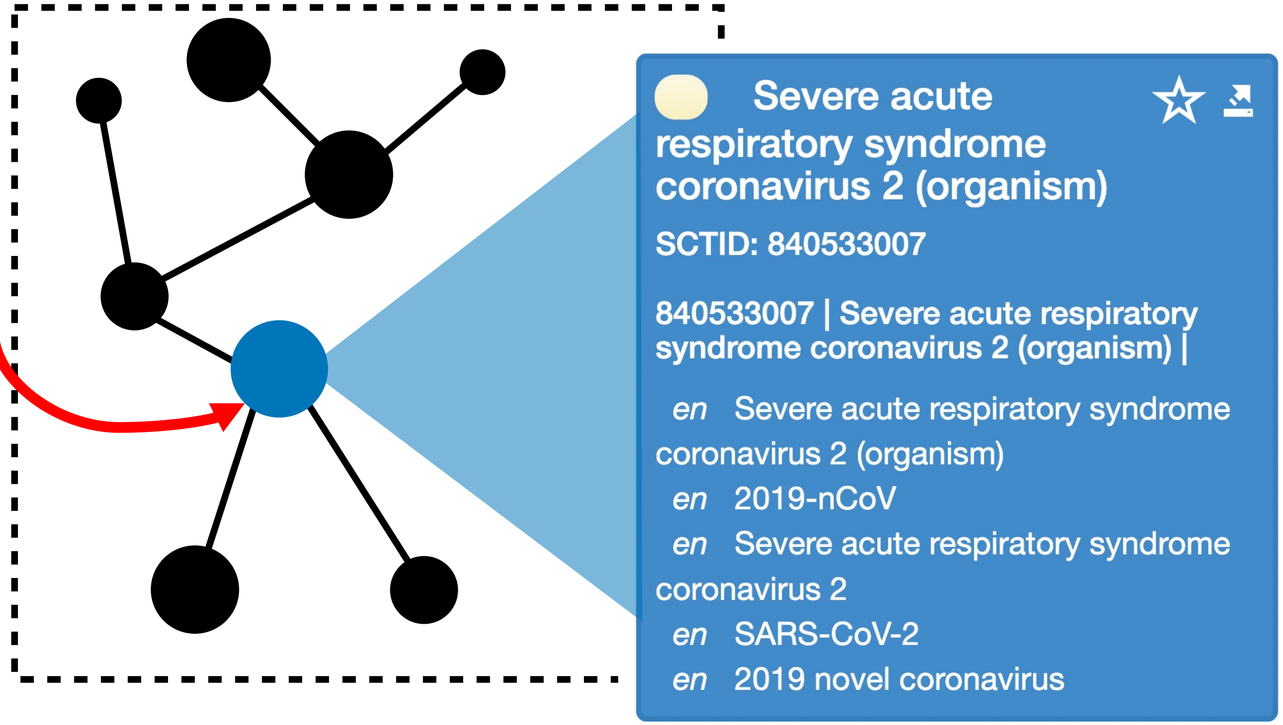
5 Evaluations

(2) Task evaluations: Medical Entity Linking

a medical **entity mention** in *free text*

OMG have u heard that John got the Covid virus ?!

entity linking



a unique concept (**node**) in a biomedical Knowledge Graph

5 Evaluations

The datasets used:

- NCBI-disease
- BC5CDR-disease
- BC5CDR-chemical
- MedMentions

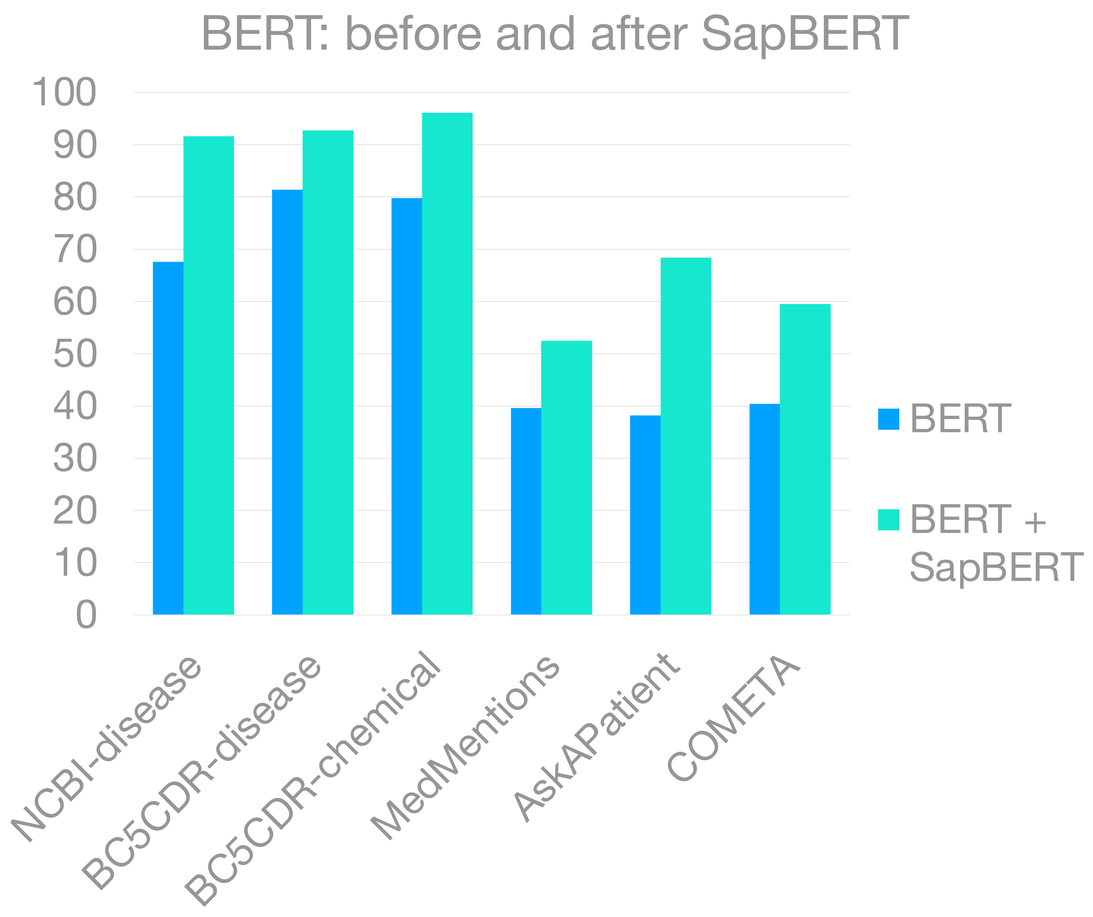
} Scientific

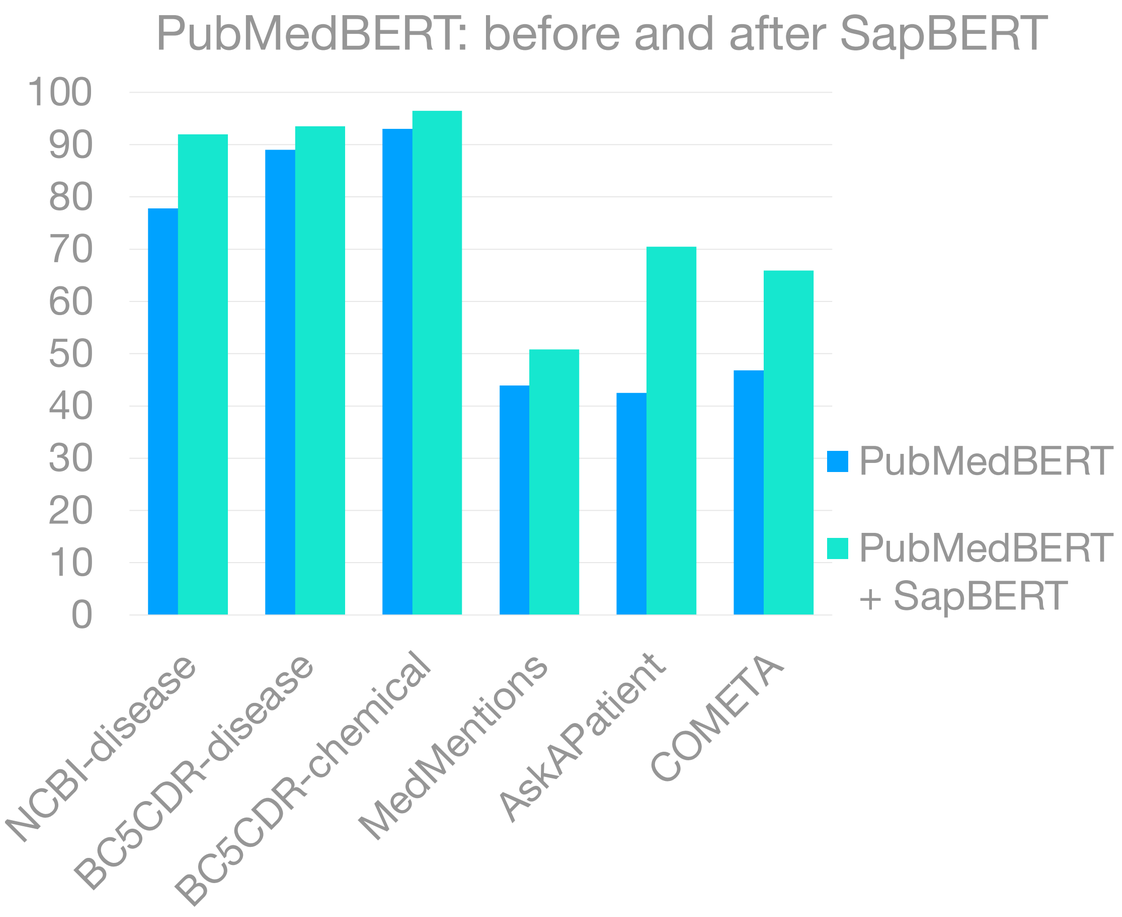
- AskAPatient
- COMETA

} Social media

5 Evaluations

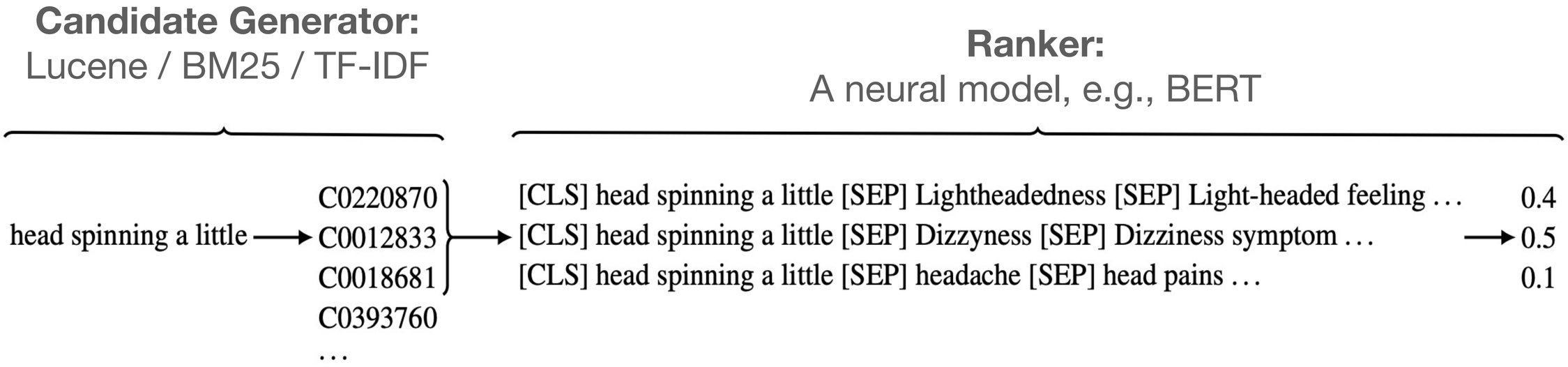
Compare with off-the-shelf pretrained Masked Language Models





5 Evaluations

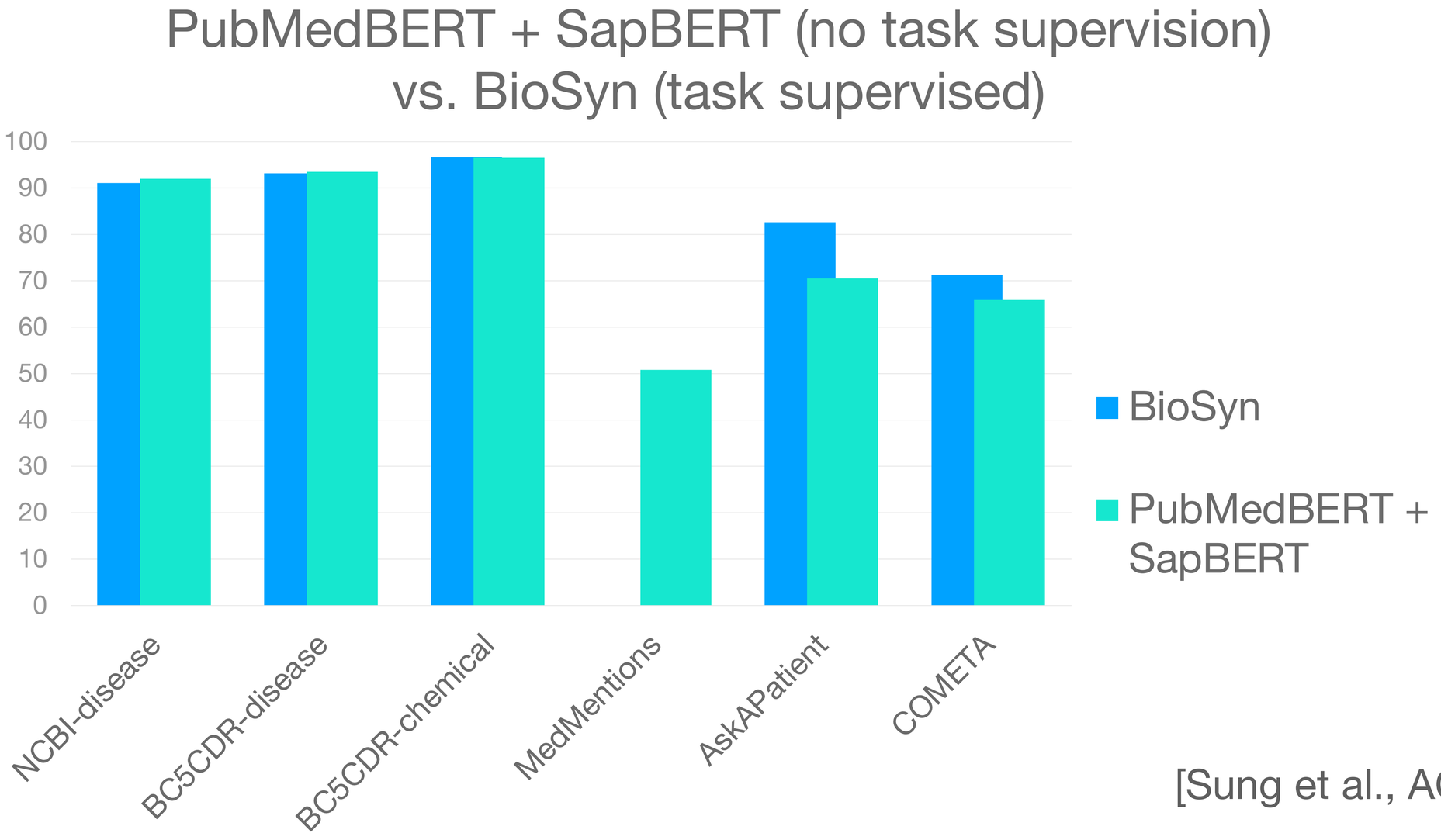
Compare with state-of-the-art biomedical entity linking systems



[Xu et al., ACL 2020] & [Sung et al., ACL 2020]

5 Evaluations

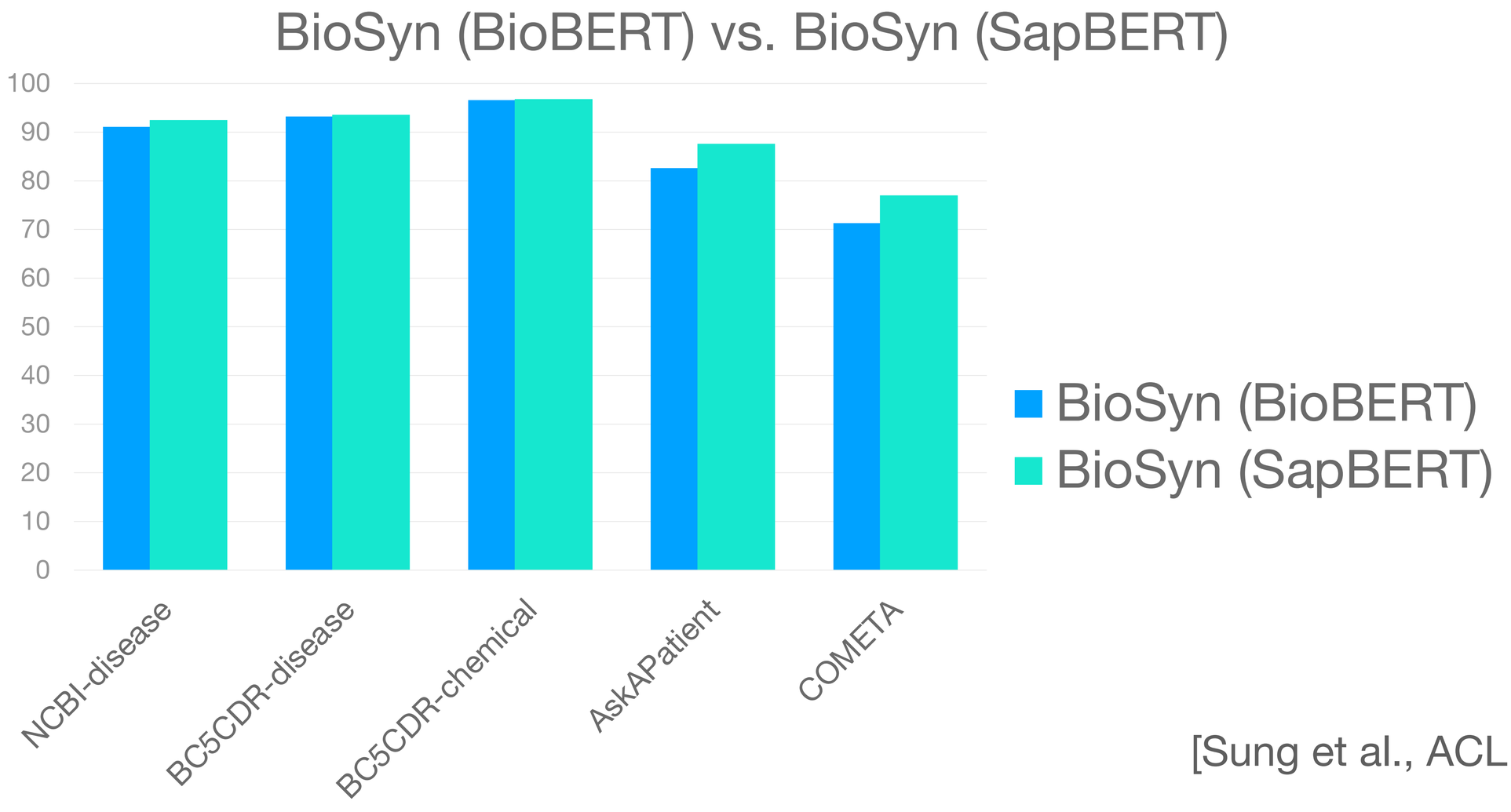
Compare with state-of-the-art biomedical entity linking systems



[Sung et al., ACL 2020]

5 Evaluations

Compare with state-of-the-art biomedical entity linking systems



[Sung et al., ACL 2020]

(2) Task evaluations: Medical Entity Linking (case studies)

[Example]* *I had **cortisone shots** when I was first diagnosed and those helped until the plaquenil built up enough to take away most of the pain.*

[Challenge] Colloquialism: *shot* -> *injection*

- Baseline model* prediction:

*- Cortisone (substance)***

- SapBERT predictions:

- Injection of cortisone (✓)
- Cortisone injection given
- Cortilymph

* All examples and baseline models are taken from the COMETA paper (EMNLP 2020). It's a string matching + BioBERT model.
** The baseline model is a hybrid of neural network and string matching, can thus only produce one prediction instead of a ranking list.

(2) Task evaluations: Medical Entity Linking (case studies)

[Example] *Get Plan B or the **Copper IUD**, if concerned about pregnancy.*

[Challenge] Abbreviation: *IUD* -> *intrauterine device*

- Baseline model prediction:

- *Cuprous oxide*

- SapBERT predictions:

- *Copper-containing intrauterine device (✓)*
- *Copper-containing intrauterine device (product) (✓)*
- *Uses copper intrauterine device contraception (finding)*

(2) Task evaluations: Medical Entity Linking (case studies)

[Example] *Can **A1c tests** be trusted to diagnose diabetes?*

[Challenge] Terminology resolution: *A1c tests* -> *glycerated hemoglobin*

- Baseline model prediction:

- Anticomplement immunofluorescence test (procedure)

- SapBERT predictions:

- Haemoglobin A1c measurement (✓)
- Hemoglobin A1c measurement (✓)
- Hemoglobin A1c measurement (procedure) (✓)

(2) Task evaluations: Medical Entity Linking (case studies)

[Example] *If you can usually sleep off migraine,* ***Benadryl*** *may help but talk to your doctor first obviously.*

[Challenge] Drug brand name

- Baseline model prediction:

- Bengal

- SapBERT predictions:

- Diphenhydramine-containing product (✓)
- Diphenhydramine (✓)
- Product containing diphenhydramine (medicinal product) (✓)

(2) Task evaluations: Medical Entity Linking (case studies)

[Example] *While there are no guarantees with **CAR**-T, and cytokine release syndrome is a real scary dangerous side effect, in your shoes I' d make it your goal to do what it takes to survive till you can get into one.*

[Challenge] Abbreviation

- Baseline model prediction:

  - Car

- SapBERT predictions:

  - Car
  - Has a car
  - CRAT

Correct answer: ***Chimeric antigen receptor***
(Future work: Context modelling)

(2) Task evaluations: Medical Entity Linking (case studies)

[Example] *I do mainly bacon and eggs now, because NO **CARBS**, delicious, and you stay full longer.*

[Challenge] Shortened form/colloquium

- Baseline model prediction:

  - Crabs

- SapBERT predictions:

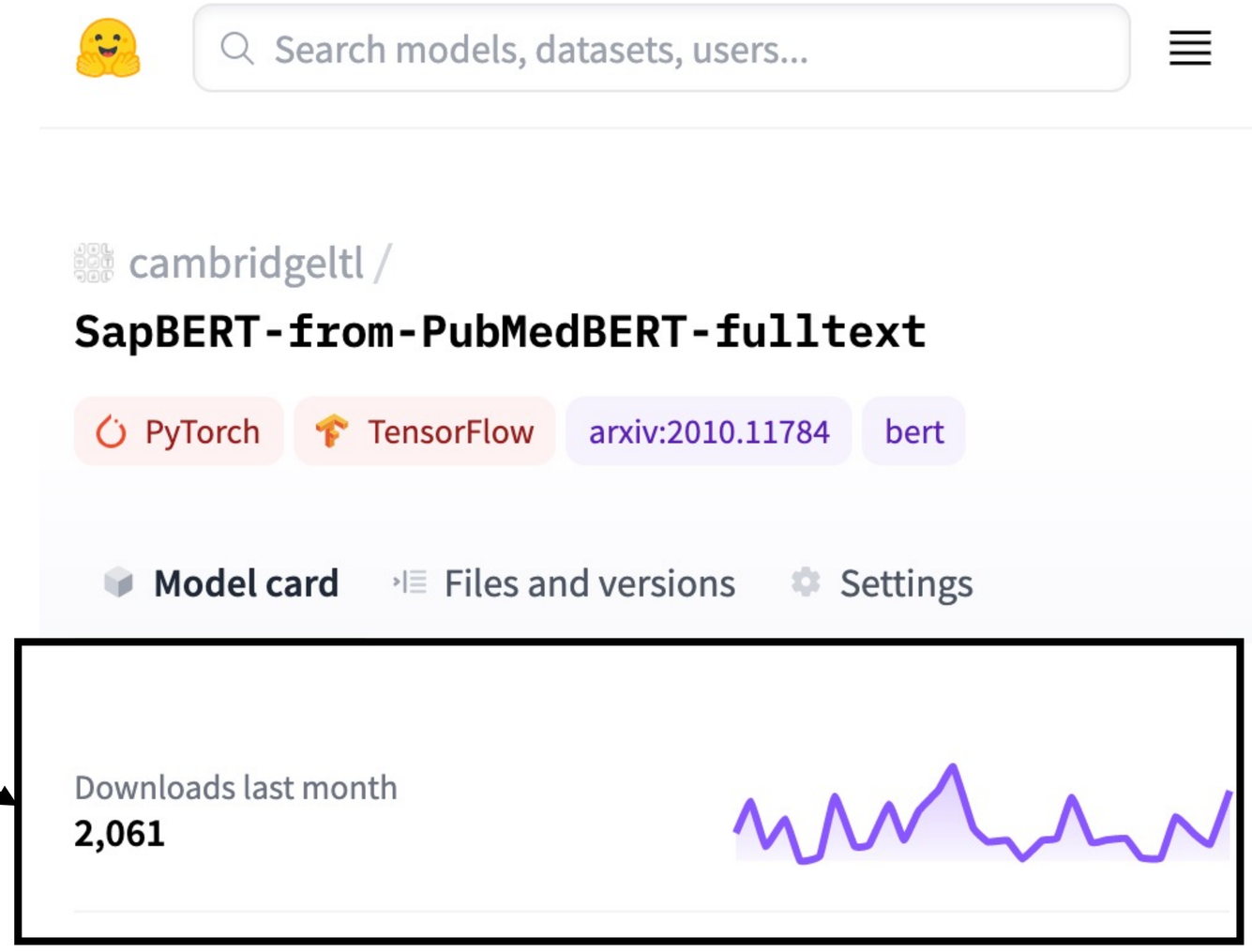
  - Carba mix
  - Carbine, device (physical object)
  - Carbine, device

Correct answer: ***Carbohydrate***
(Some layman language can not be learned from standardized vocabularies.)

6 Open-Source

SapBERT is receiving positive feedbacks and gaining popularity within the Machine Learning and Natural Language Processing communities

The SapBERT model gets 2,000+ downloads per month

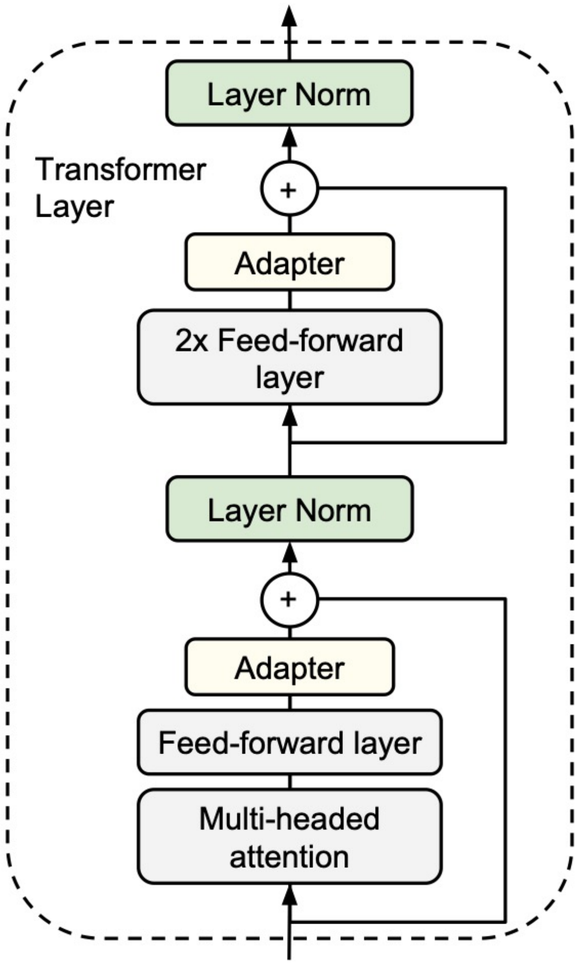


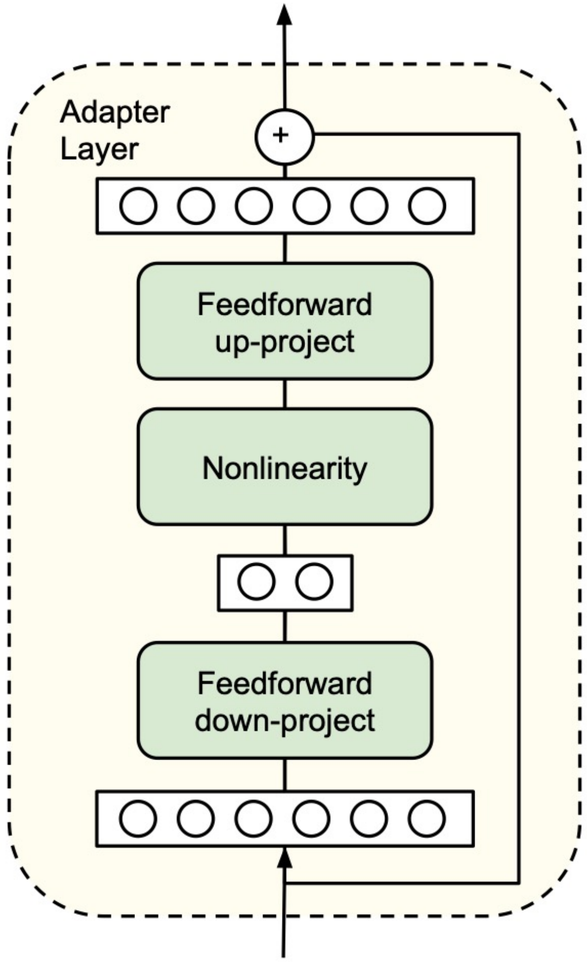
7 Work-in-Progress

7 Work-in-Progress

How about sentence-level tasks? (QA, text classification, etc.)

Adapters!

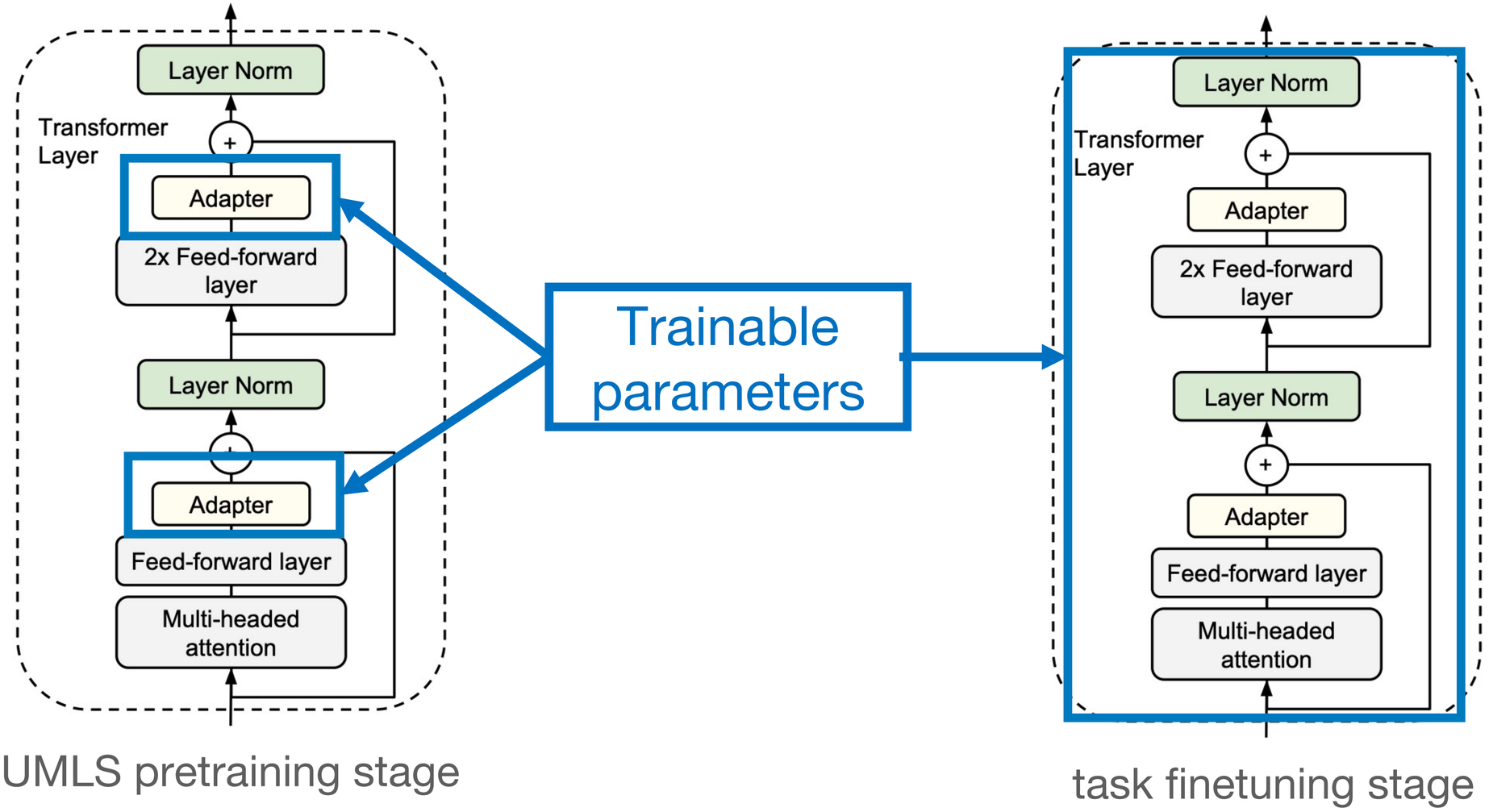




[Houlsby et al., ICML 2019]

7 Work-in-Progress

Adapter UMLS pretraining and task finetuning



UMLS pretraining stage

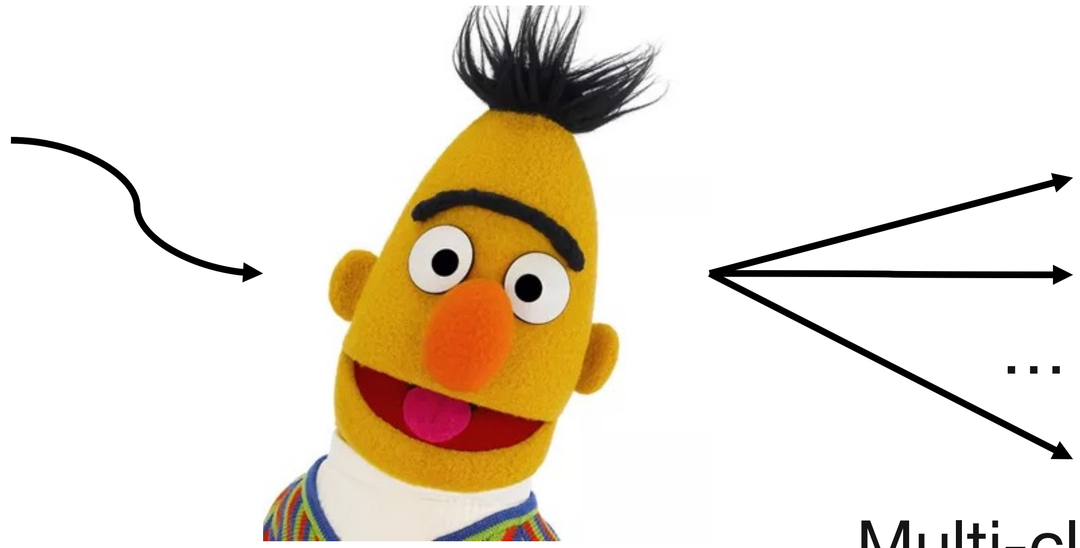
task finetuning stage

7 Work-in-Progress

We train a knowledge graph completion objective on UMLS triples:

(chemical_A, *interact_with*, ?)

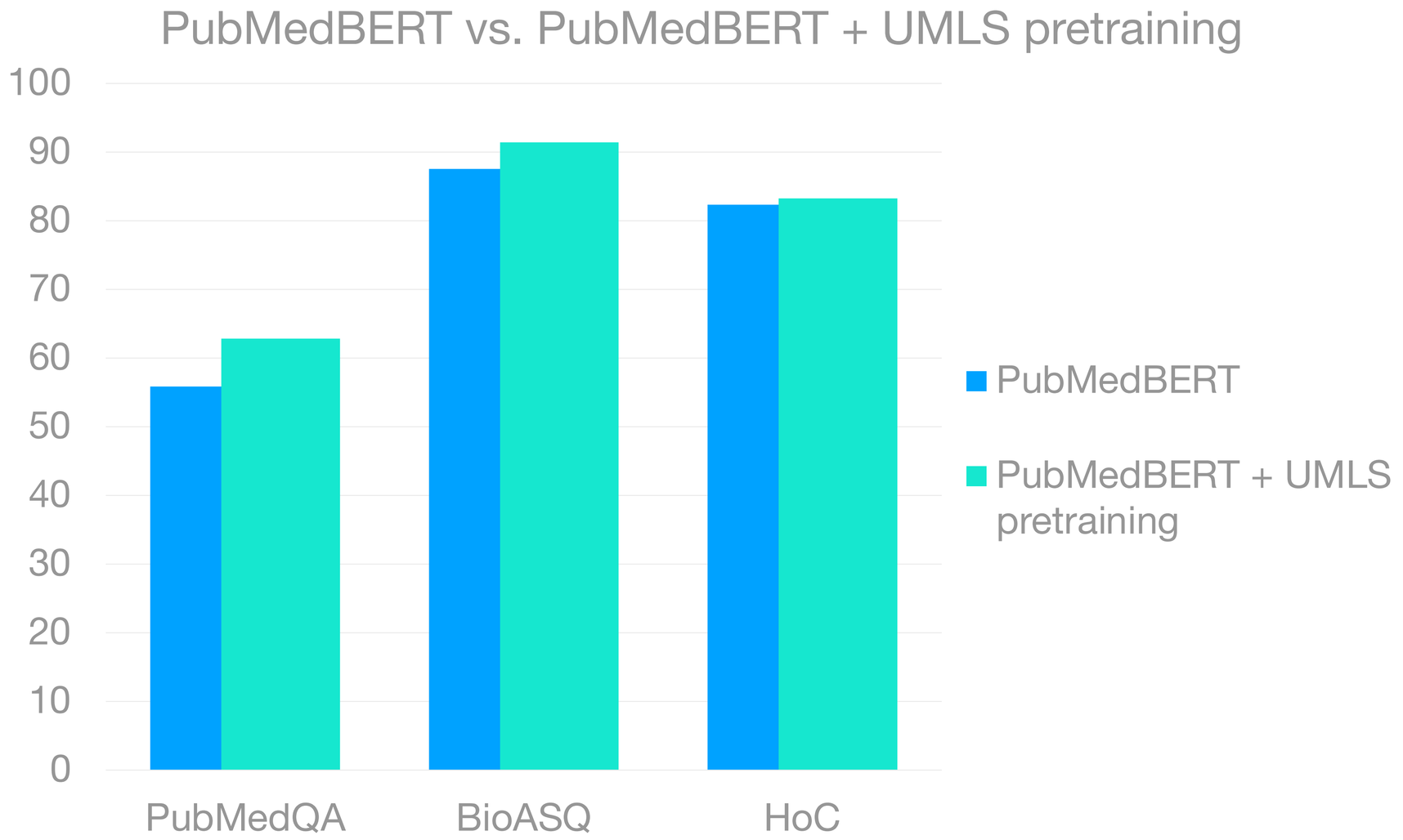
Input: [CLS] chemical_A [SEP] interact [SEP]



Multi-class classification

7 Work-in-Progress

Sentence-level tasks results



Thank you!

fl399@cam.ac.uk