UNIVERSITY OF CAMBRIDGE

Automated coding of patient-centric entities using neural natural language processing

12th March 2021

Nigel Collier
Theoretical and Applied Linguistics, MMLL

Language Technology Lab

* Thanks to

At the LTL Cambridge:

- Marco Basaldella
- Fangyu Liu
- Ehsan Shareghi
- Zaiqiao Meng

Supported by

EPSRC
Engineering and Physical Sciences Research Council

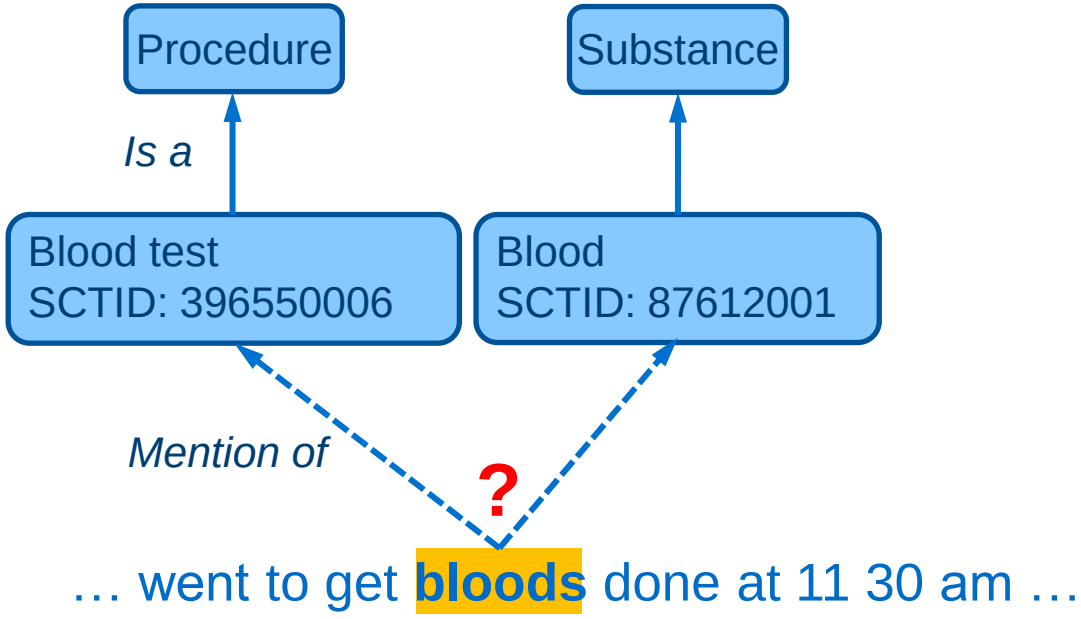
MRC Medical Research Council

HDRUK
Health Data Research UK

OVERVIEW

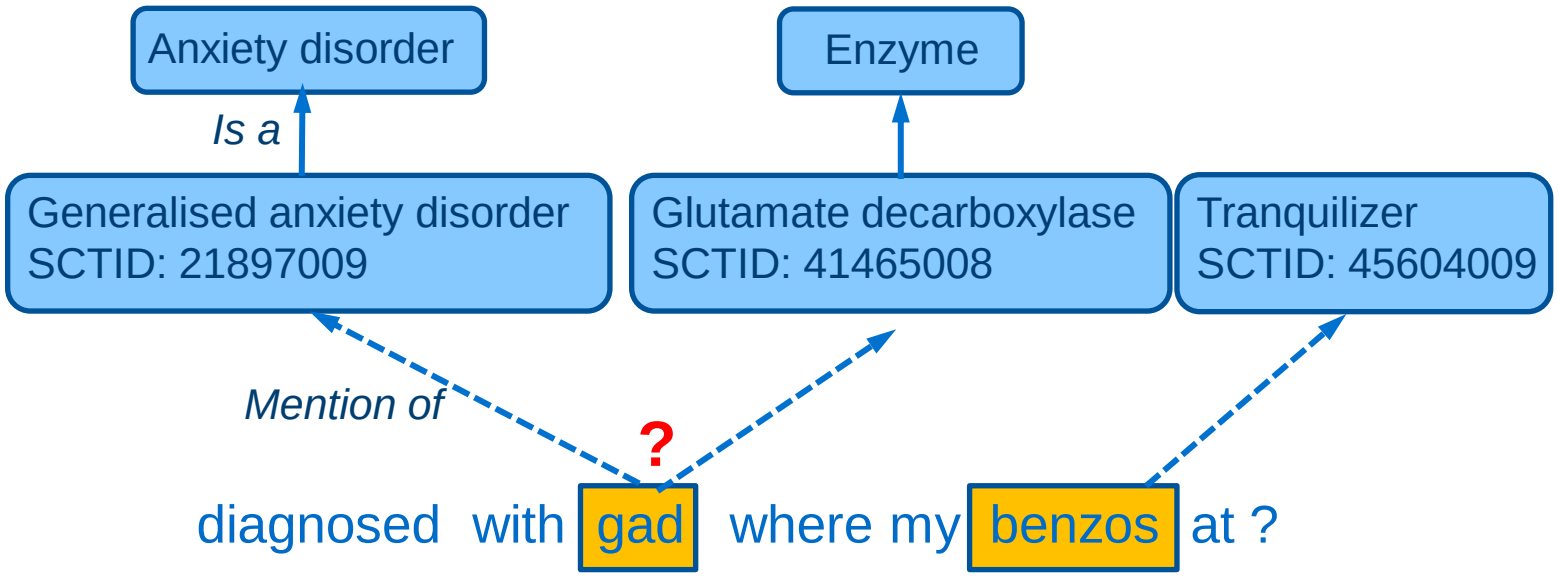
Entity Coding: a Central Task in Text Mining

Coding patient vocabulary to SNOMED CT



Entity Coding: a Central Task in Text Mining

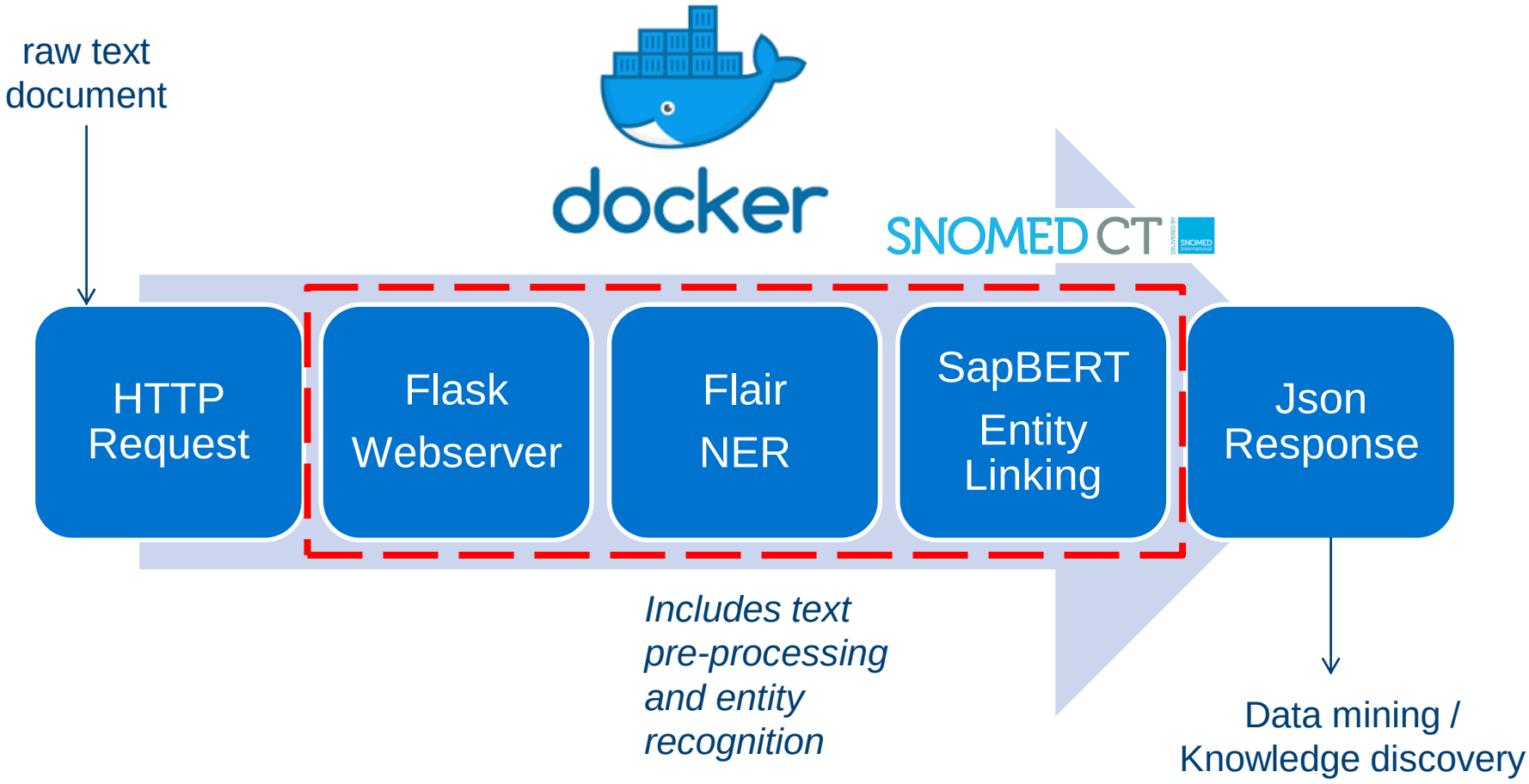
Coding patient vocabulary to SNOMED CT



Illustrating the Complexities of Entity Coding in Health

Source	Entity Mention	Target Concept (SNOMED)
Twitter	hungry	hunger
Twitter	gained 2kgs in weight	weight gain
Twitter	head spinning	dizziness
Twitter	ruptured his bowel	gastrointestinal perforation
EHR	No pneumothorax	history of pneumothorax, negative
EHR	right breast cancer	breast cancer + right
EHR	A.FIB	atrial fibrillation
EHR	tumor … in left ovary	tumor of ovary
Literature	thrombophilic condition	blood coagulation disorder [not thrombophilia]
Literature	peculiar changes in the dendrites of Purjinje cells	abnormal + Purjinje cell + dendrite + associated morphology

HELIN (Health Entity Linking) pipeline



Example API output

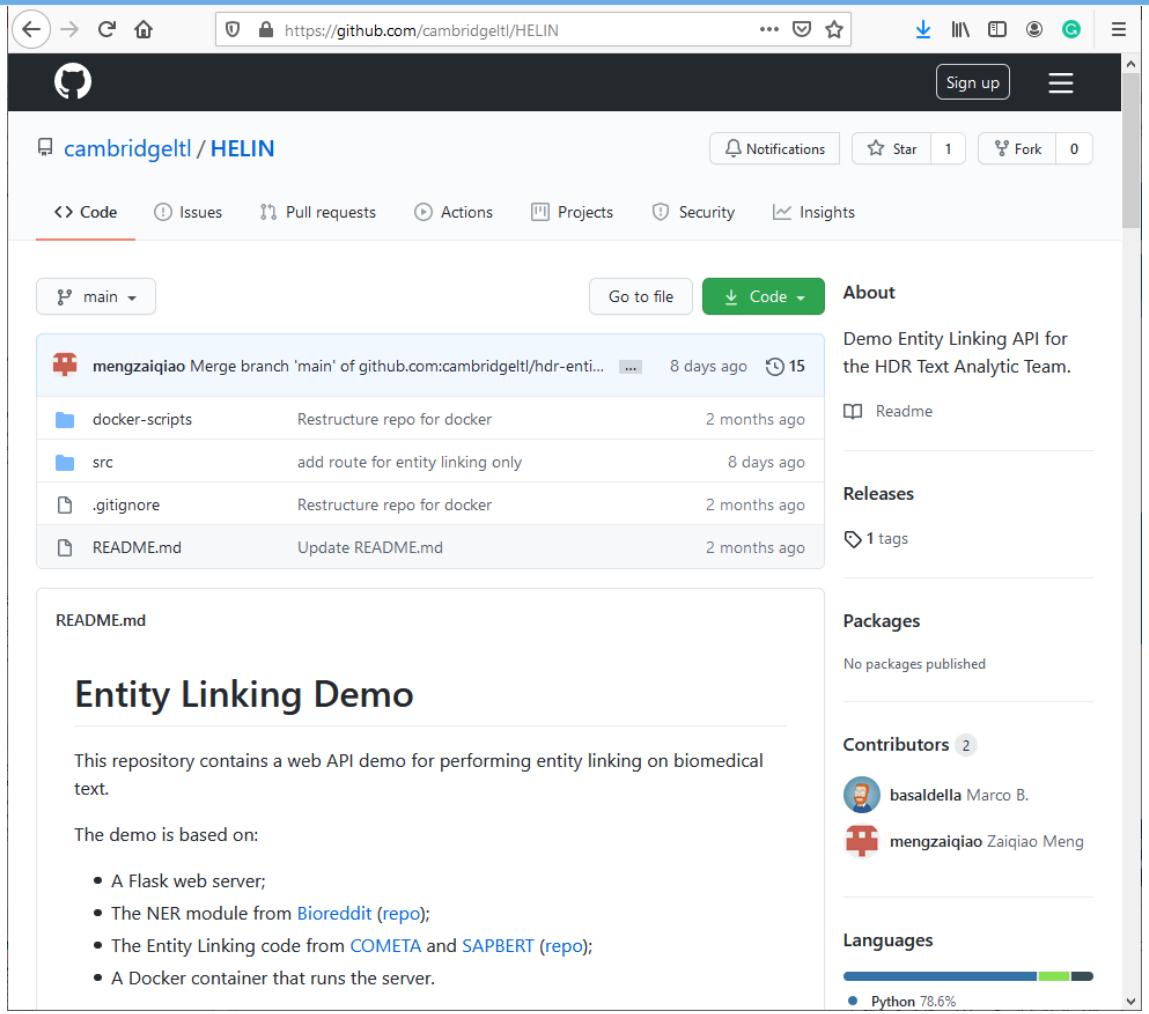
API call:

- `/tag_string?txt='I woke up with migraine so I took an aspirine'`
- Sentence: *I woke up with migraine so I took an aspirine*
- Correctly detected both entities
- Resolved both to SNOMED *even if* Aspirin is misspelled


```
entities:
  0:
    0:      "T1"
    1:      "Phenotype"
    2:
      0:
        0:  16
        1:  24
    3:      "Migraine"
    4:      "SCTID: 37796009"
  1:
    0:      "T2"
    1:      "Molecule"
    2:
      0:
        0:  38
        1:  46
    3:      "Aspirin"
    4:      "SCTID: 387458008"
text:       "'I woke up with migraine so I took an aspirine'"
```

Where can I get HELIN?

https://github.com/cambridgeltl/HELIN



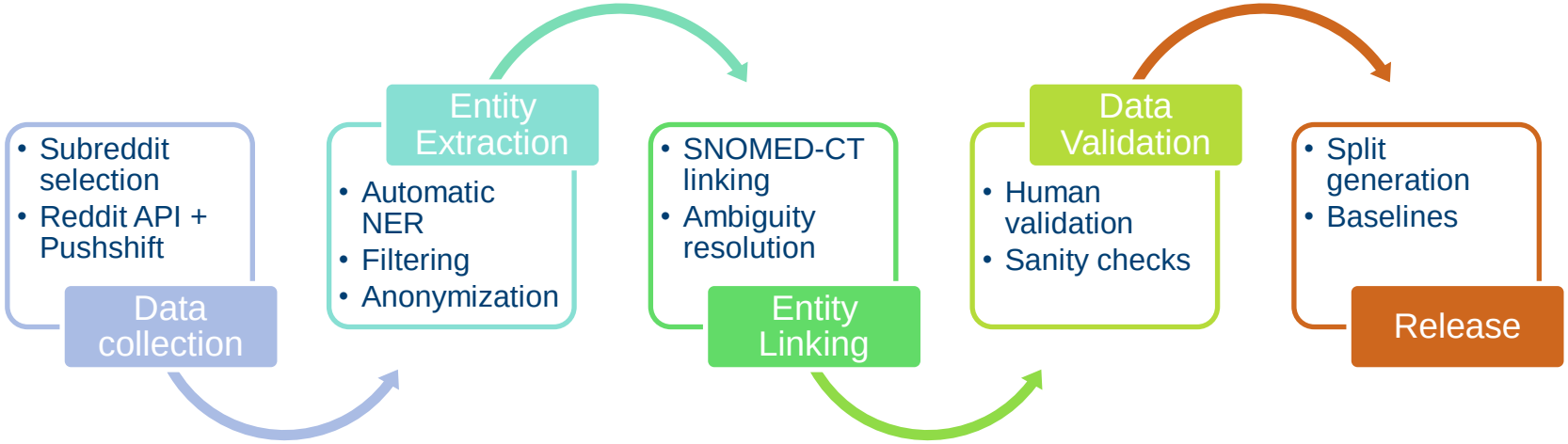
TECHNICAL DETAILS: DATA

[1] Basaldella, M., & Collier, N. (2019, November). BioReddit: Word embeddings for user-generated biomedical NLP. In Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019) (pp. 34-38).
[2] Liu, F., Shareghi, E., Meng, Z., Basaldella, M. and Collier, N. Self-alignment Pre-training for Biomedical Entity Representations. In Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2021), Mexico City, Mexico), in press.
[3] Basaldella, M., Liu, F., Shareghi, E., & Collier, N. (2020, November). COMETA: A Corpus for Medical Entity Linking in the Social Media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3122-3137).

COMETA: a training set from Reddit data

- Timeframe: 2015 to 2018
- Theme: 68 subreddits
  - General domain: `r/AskDocs`, `r/DiagnoseMe`, `r/health`, `r/AskAPharmacist`, `r/AskADentist`, `r/HealthInsurance`, ...
  - Specific issues: `r/flu`, `r/obgyn`, `r/cancer`, `r/diabetes`, `r/migraine`, `r/benzorecovery`, ...
- **User anonymization plus pseudo-anonymization**
- **Avoid** low-traffic & links-only subreddits
- **Removed** automated/bot posts

Constructing COMETA



[3] Basaldella, M., Liu, F., Shareghi, E., & Collier, N. (2020, November). COMETA: A Corpus for Medical Entity Linking in the Social Media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3122-3137).

COMETA: examples

Input term	right ankle
Context	I have had an avulsion fracture lately in my right ankle
Target general SNOMED label	Structure of right ankle (STCID: 6685009)
Target specific SNOMED label	Structure of right ankle (STCID: 6685009)

COMETA: examples

Input term	bloods
Context	Went to get bloods done at NUM and results came back as fine
Target general SNOMED label	Blood (STCID: 87612001)
Target specific SNOMED label	Blood test (STCID: 396550006)

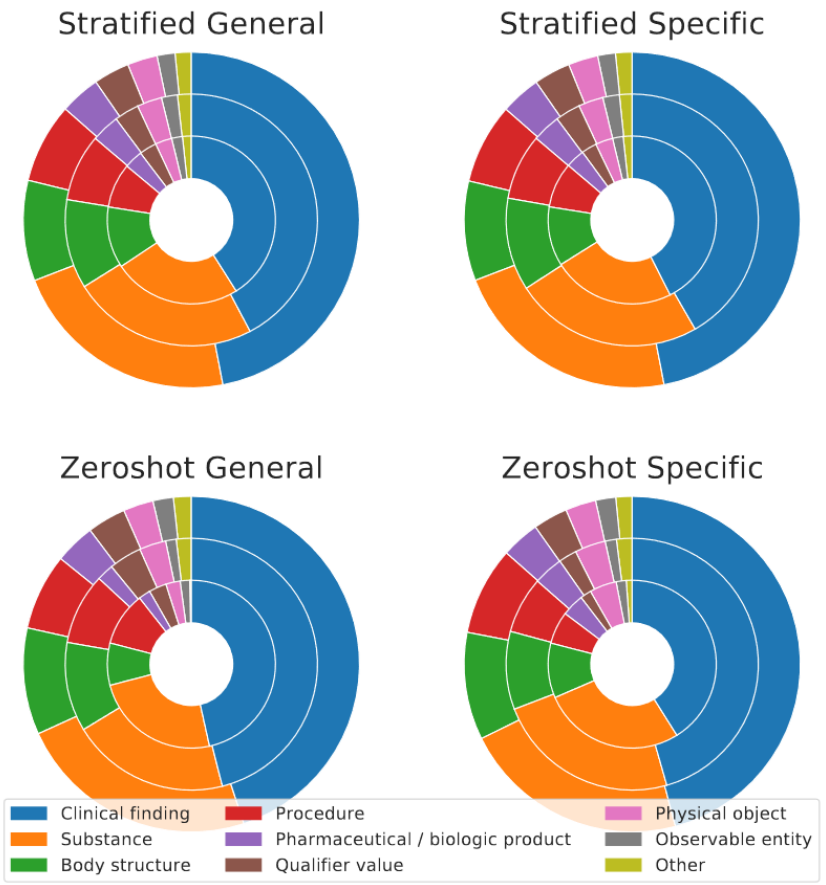
COMETA: examples

Input term	CPAP
Context	I remember years ago getting CPAP supplies and realizing for the first time what a crock it was
Target general SNOMED label	Continuous positive airway pressure ventilation treatment (STCID: 47545007)
Target specific SNOMED label	Continuous positive airway pressure ventilation treatment (STCID: 47545007)

COMETA splits: encouraging zero shot concept recognition

Split		Training	Dev	Test
Stratified	**General**	13489	2176	4350
	Specific	13441	2205	4369
Zero-Shot	**General**	14062	1958	3995
	Specific	13714	2018	4283

* Number of sentences with one target concept per sentence



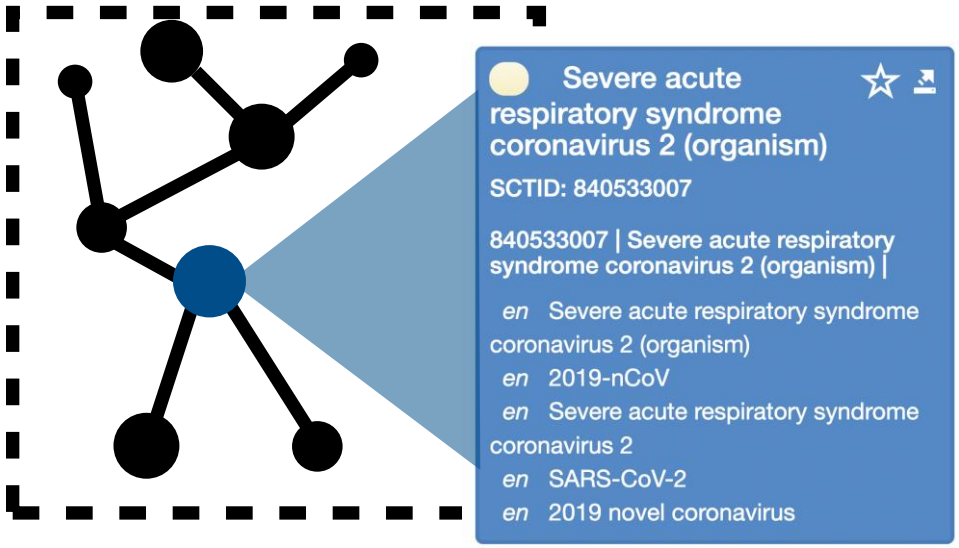
TECHNICAL DETAILS: SAPBERT

[1] Basaldella, M., & Collier, N. (2019, November). BioReddit: Word embeddings for user-generated biomedical NLP. In Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019) (pp. 34-38).
[2] Liu, F., Shareghi, E., Meng, Z., Basaldella, M. and Collier, N. Self-alignment Pre-training for Biomedical Entity Representations. In Proceedings of the 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2021), Mexico City, Mexico), in press.
[3] Basaldella, M., Liu, F., Shareghi, E., & Collier, N. (2020, November). COMETA: A Corpus for Medical Entity Linking in the Social Media. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 3122-3137).

Magic sauce: 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

Self-Alignment Pretraining to recognize synonyms

Technical challenge: UMLS is huge, but not always informative

- Concept: [C0020336] hydroxychloroquine
- Semantic Types
- Definitions
- Atoms (44) string [AUI / RSAB / TTY / Code]
  - hydroxychloroquine [A22730801/ATC/IN/P01BA02]
  - hydroxychloroquine [A18610592/CHV/PT/0000006376]
  - hydroxychloroquine [A0481254/CSP/ET/2530-4570]
  - (±)-hydroxychloroquine [A27060116/DRUGBANK/SY/DB01611]
  - 2-((4-((7-chloro-4-quinolyl)amino)pentyl)ethylamino)ethanol [A3013911
  - 2-(N-(4-(7-chlor-4-chinolylamino)-4-methylbutyl)ethylamino)ethanol [A3
  - 7-chloro-4-(4-(ethyl(2-hydroxyethyl)amino)-1-methylbutylamino)quinolir
  - 7-chloro-4-(4-(N-ethyl-N-β-hydroxyethylamino)-1-methylbutylamino)qui
  - 7-chloro-4-[4-(N-ethyl-N-β-hydroxyethylamino)-1-methylbutylamino]quir
  - 7-chloro-4-[5-(N-ethyl-N-2-hydroxyethylamino)-2-pentyl]aminoquinoline
  - Hidroxicloroquina [A30138425/DRUGBANK/FSY/DB01611]
  - Hydroxychloroquine [A27058292/DRUGBANK/IN/DB01611]
  - Hydroxychloroquinum [A30136311/DRUGBANK/FSY/DB01611]

.....

synonyms of *hydroxychloroquine*

Most of *hydroxychloroquine*'s variants are easy:

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

But a few can be very hard:

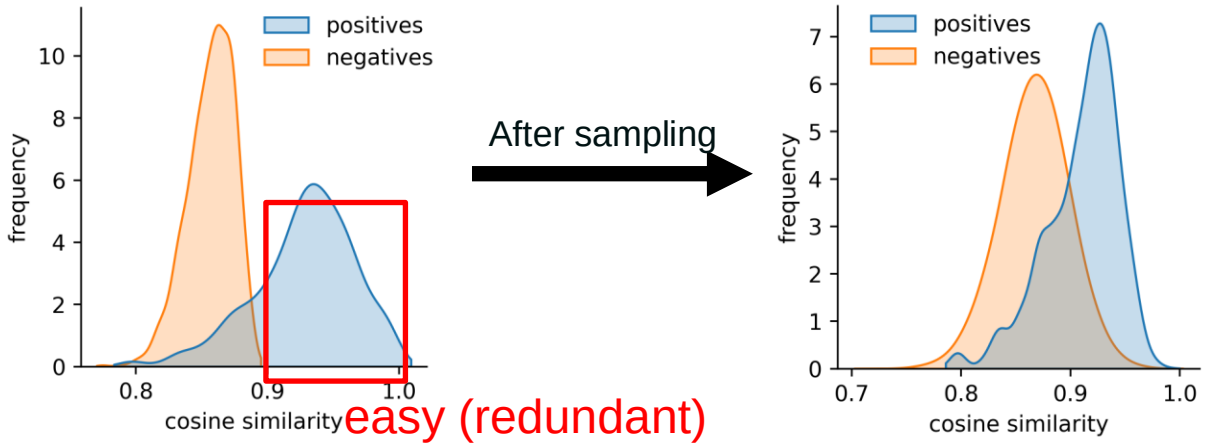
- *HCQ*
- *plaquenil*
- ……

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

Self-Alignment Pretraining with harder examples

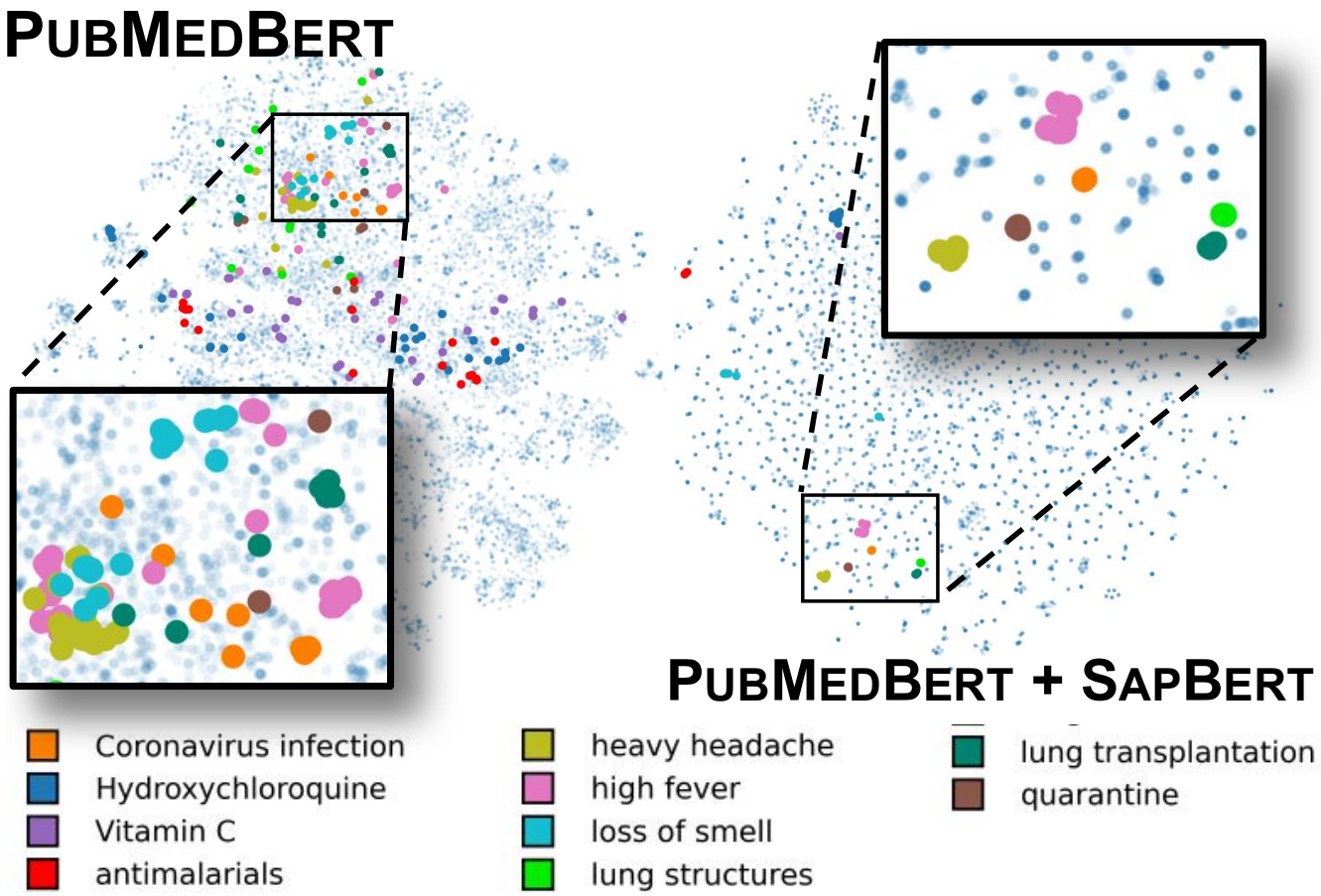
Techniques: (1) smart online sampling (2) multi-similarity loss

- (1) **smart online sampling**: drop the easily-solvable examples



- (2) **multi-similarity loss**: learn more from the more informative examples

A qualitative evaluation using T-SNE visualisation



Quantitative evaluations using COMETA and other benchmarks

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	96.6	OOM	87.5	79.0
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
+ SAPBERT (FINE-TUNED)	92.3	93.2	96.5	50.4	**89.0**	**81.1**
BIOSYN	91.1	93.2	96.6	OOM	82.6	71.3
+ (init. w/) SAPBERT	**92.5**	**93.6**	**96.8**	OOM	87.6	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 underline denote the best and second best results in the column.

Huge performance boost when applied to popular Masked-Language Models

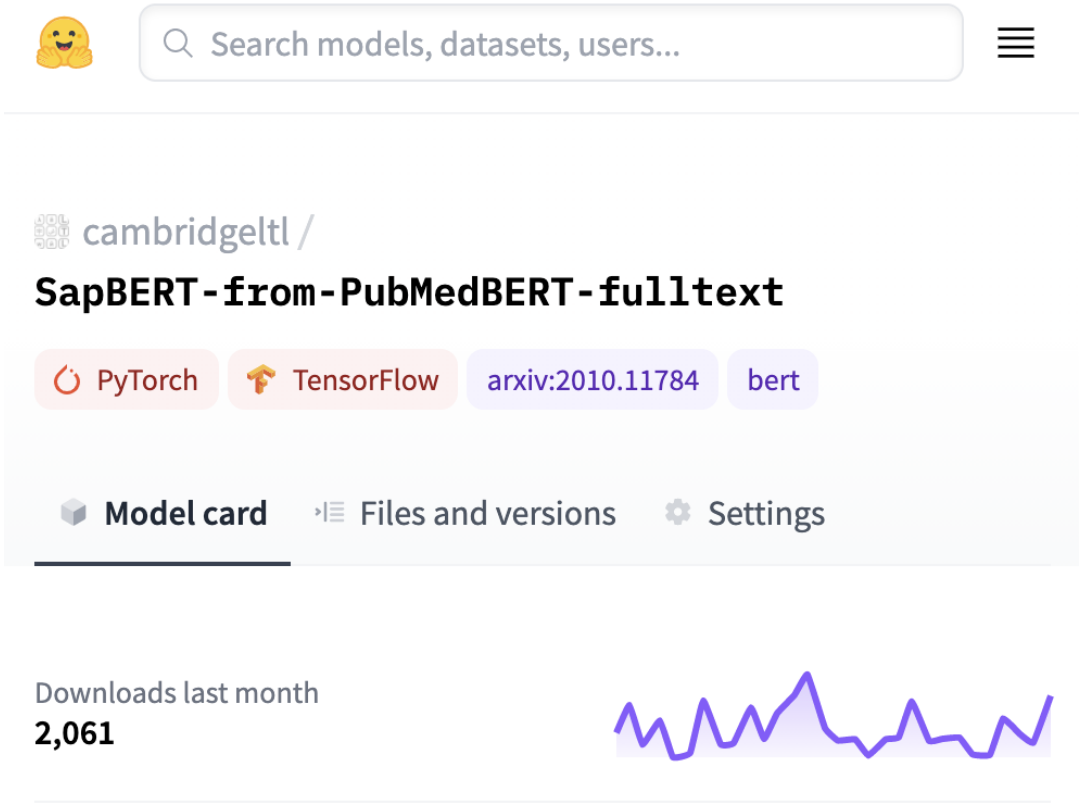
State-of-the-art results across 6 academic benchmarks

Accuracy @1 results on D1: NCBI, D2: BC5CDR-d, D3: BC5CDR-c, D4: MedMentions, D5: AskAPatient, D6: COMETA

SapBERT is available via Github and HuggingFace

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



Thank you!

https://sites.google.com/site/nhcollier/

nhc30@cam.ac.uk

ORCID: 0000-0002-7230-4164

Twitter: @nigelhcollier