SapBERT: Self-Alignment Pretraining for Biomedical Entity Representations (Accepted to NAACL 2021.)



UNIVERSITY OF CAMBRIDGE

Language Technology Lab, University of Cambridge, UK Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, Nigel Collier

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



positive (synonyms of the anchor)

negative (non-synonyms of the anchor)

4 Method: Self-Alignment Pretraining **Technical challenge: UMLS is gigantic, but not always informative**

- Concept: [C0020336] hydroxychloroquine
- Semantic Types

synonyms of *hydroxychloroquine*

- Definitions
- string [AUI / RSAB / TTY / Code] Atoms (44)
 - hydroxychloroquine [A22730801/ATC/IN/P01BA02]
 - hydroxychloroquine [A18610592/CHV/PT/000006376]
 - 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]

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?



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



4 Method: Self-Alignment Pretraining Techniques: (1) smart online sampling



4 Method: Self-Alignment Pretraining Techniques: (2) multi-similarity loss



push away negatives

$$\left(1 + \sum_{n \in \mathcal{N}_i} e^{\alpha(\mathbf{S}_{in} - \epsilon)} \right)$$
$$+ \sum_{p \in \mathcal{P}_i} e^{-\beta(\mathbf{S}_{ip} - \epsilon)} \right) ,$$

cluster positives

5 Evaluations (1) **T-SNE** visualisations

PUBMEDBERT





Coronavirus infection Hydroxychloroquine Vitamin C antimalarials



PUBMEDBERT + SAPBERT

heavy headache high fever loss of smell lung structures



Iung transplantation quarantine

5 Evaluations (2) Task evaluations: Medical Entity Linking

entity linking

a medical <u>entity mention</u> in *free text*

OMG have u heard that



a unique concept (node) in a biomedical Knowledge Graph



☆ 🛎 respiratory syndrome coronavirus 2 (organism) SCTID: 840533007 840533007 | Severe acute respiratory syndrome coronavirus 2 (organism) en Severe acute respiratory syndrome coronavirus 2 (organism) en 2019-nCoV en Severe acute respiratory syndrome en SARS-CoV-2 en 2019 novel coronavirus



The datasets used:

NCBI-disease
BC5CDR-disease
BC5CDR-chemical
MedMentions

- AskAPatient - COMETA Scientific

Social media

Compare with off-the-shelf pretrained Masked Language Models





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

Candidate Generator: Lucene / BM25 / TF-IDF

...

[CLS] head spinning a little [SEP] Lightheadedness [SEP] Light-headed feeling ... C0220870 0.4 ► [CLS] head spinning a little [SEP] Dizzyness [SEP] Dizziness symptom ... head spinning a little \longrightarrow C0012833 **→**0.5 [CLS] head spinning a little [SEP] headache [SEP] head pains ... 0.1 C0018681 C0393760

Ranker: A neural model, e.g., BERT

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



Compare with state-of-the-art biomedical entity linking systems PubMedBERT + SapBERT (no task supervision) vs. BioSyn (task supervised)



[Sung et al., ACL 2020]



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



[Sung et al., ACL 2020]



[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)**

* 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.

• SapBERT predictions:

- Injection of cortisone (\checkmark) - Cortisone injection given - Cortilymph

[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 (\checkmark) - Copper-containing intrauterine device (product) (\checkmark) - Uses copper intrauterine device contraception (finding)



[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) (1)

[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

- product) (\checkmark)

• SapBERT predictions:

- Diphenhydramine-containing product (\checkmark) - Diphenhydramine (\checkmark) - Product containing diphenhydramine (medicinal

[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 - Car - Has a car - CRAT

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

• SapBERT predictions:

[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 - Carba mix

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

• SapBERT predictions:

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

6 Open-Source

the Machine Learning and Natural Language Processing communities



SapBERT is receiving positive feedbacks and gaining popularity within

7 Work-in-Progress

7 Work-in-Progress

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





[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]





7 Work-in-Progress **Sentence-level tasks results**



Thank you!

fl399@cam.ac.uk