Self-alignment Pre-training for Biomedical Entity Representations

Fangyu Liu[†], **Ehsan Shareghi**[†,‡], **Zaiqiao Meng**[†], **Marco Basaldella**[†], **Nigel Collier**[†]
[†]Language Technology Lab, University of Cambridge
[‡]University College of London
[†]{fl399, zm324, mb2313, nhc30}@cam.ac.uk
[‡]e.shareghi@ucl.ac.uk

Abstract

Despite the widespread success of self-supervised learning via masked language models, learning representations directly from text to accurately capture complex and fine-grained semantic relationships in the biomedical domain remains as a challenge. Addressing this is of paramount importance for tasks such as entity linking where complex relational knowledge is pivotal. We propose SAPBERT, a pre-training scheme based on BERT. It self-aligns the representation space of biomedical entities with a metric learning objective function leveraging UMLS, a collection of biomedical ontologies with >4M concepts. Our experimental results on six medical entity linking benchmarking datasets demonstrate that SAPBERT outperforms many domain-specific BERT-based variants such as BIOBERT, BLUEBERT and PUBMEDBERT, achieving the state-of-the-art (SOTA) performances.

1 Introduction

Biomedical entity[1] representation is the foundation for a plethora of text mining systems in the medical domain, facilitating applications such as literature search [LKL+16], clinical decision making [RSVH15] and relational knowledge discovery (e.g. chemical-disease, drug-drug and protein-protein relations) [WLA+18]. The heterogeneous naming of biomedical concepts pose a major challenge to representation learning. For example, the medication *Hydroxychloroquine* (CUI: C0020336) is often referred to as *Oxichlorochine* (alternative name), *HCQ* (social media) and *Plaquenil* (brand name). Medical entity linking is a segue to tackle this problem by framing it as a task of mapping entity mentions to unified concepts in a medical knowledge graph. However, the quality of the learned representations remains as the bottleneck for medical entity linking system [BLSC20]. Prior works in this domain have adopted very sophisticated text pre-processing heuristics [JWX20, SJLK20] which can hardly cover all the variations of biomedical names. Beyond the biomedical domain, entity linking has been intensively investigated on the less challenging general domain [SWH15, KGH18].

In parallel, self-supervised learning (SSL) has shown tremendous success in visual representation learning with performance of unsupervised methods approaching supervised ones [HFW+20, CKNH20, GSA+20], as well as NLP via leveraging masked language modelling (MLM) objective to learn semantics from distributional representations [DCLT19, LOG+19].[2] Domain-specific pre-training on biomedical corpora (e.g. BIOBERT [LYK+20] and BLUEBERT [PYL19]) have seen many progress in biomedical text mining tasks. Nonetheless, representing medical entities with the existing

[1] In this work, *biomedical entity* refers to the surface forms of biomedical concepts, which can be a single word (e.g. *fever*), a compound (e.g. *sars-cov-2*) or a short phrase (e.g. *abnormal retinal vascular development*).

[2] The MLM framework is similar to denoising auto-encoders trying to reconstruct part(s) of the text [KdMdY+20, LLG+20] and can be considered as a form of self-training.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

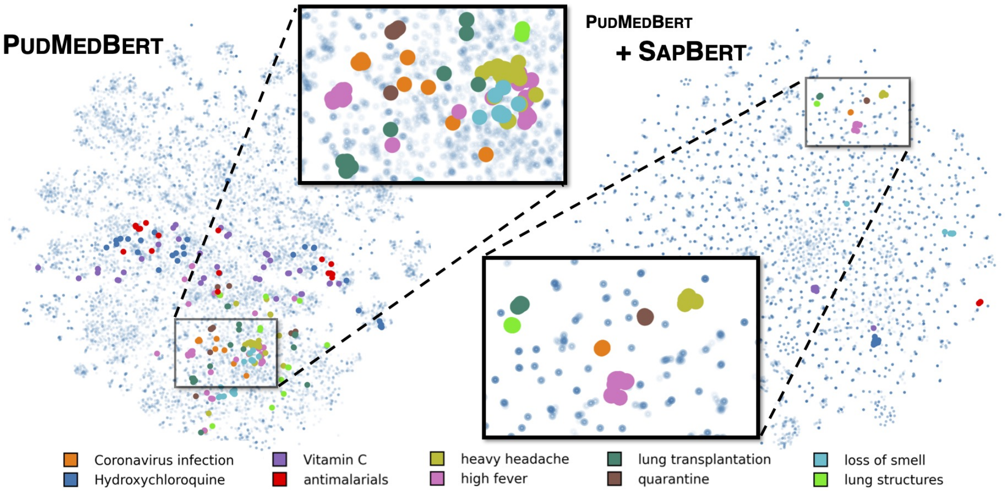

Figure 1: The t-SNE [MH08] visualisation of UMLS entities under PUBMEDBERT (BERT pre-trained on PubMed papers) & PUBMEDBERT+SAPBERT (PUBMEDBERT further pre-trained on UMLS synonyms). Without self-alignment pre-training, biomedical names of different concepts are hard to separate from each other in the heterogeneous embedding space. After the pre-training, concepts' synonyms are self-aligned and form compact clusters where names of different concepts live in distinct locations in the embedding space.

SOTA self-trained BERT (e.g. PUBMEDBERT [GTC$^+$20]) model as suggested in Fig. 1(left) does not lead to a well-separated representation space.

To address the aforementioned issue, we propose to pre-train a Transformer-based language model (BERT) on the biomedical knowledge graph of UMLS [Bod04], which is the largest interlingua of biomedical ontologies and contains a comprehensive collection of biomedical synonyms in various forms (UMLS 2020AA has more than 4 million concepts and 10 million synonyms which stem from over 150 controlled vocabularies including MeSH, SNOMED CT, RxNorm, Gene Ontology and OMIM).[3] We pre-train BERTs on UMLS with a self-alignment objective that clusters synonyms of the same concept using a metric learning loss.

We show that with a well-crafted non-parametric metric learning formulation which scales well on UMLS, an end-to-end Transformer-based language model is sufficient to perform well on the task of medical entity linking (MEL). The MEL task aims to map a medical mention to a well-defined controlled vocabulary (usually a knowledge graph). In contract with the current approaches which adopt complex pipelines and hybrid components [XZB20, JWX20], our **S**elf-**a**ligning **p**retrained **BERT** (SAPBERT) applies a much simpler training procedure without relying on any pre- or post-processing steps, achieving the SOTA performance with a simple nearest neighbour's search. On multiple scientific language MEL benchmarks, SAPBERT outperforms previous SOTA even without any finetuning on specific MEL datasets. On social media language MEL datasets, SAPBERT outperforms SOTA after fine-tuning on the social-media-domain training set. When compared with other domain-pre-trained contextual models (e.g. BIOBERT, BLUEBERT and PUBMEDBERT), SAPBERT can outperform them by up to 20% on Accuracy across different benchmarks.

2 Method: Self-Alignment Pre-training

To address heterogeneous naming issue, we design our pre-training scheme with a self-alignment step based on a metric learning framework. This step utilises an existing BERT model and learns to align the biomedical names with their synonyms from the UMLS knowledge graph. In the following we introduce our metric learning framework that is able to conduct both the self-alignment pre-training step on UMLS synonyms set, and the fine-tuning step on task-specific datasets.

Formal Definition. Let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ denote a tuple of a name and its categorical label. For the self-alignment pre-training step, $\mathcal{X} \times \mathcal{Y}$ is the set of all (name, CUI[4]) pairs in UMLS, e.g.

[3] https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html

[4] In UMLS, CUI is the **C**oncept **U**nique **I**dentifier.

(*Remdesivir*, C4726677); while for the fine-tuning step, it is formed as an entity mention and its corresponding mapping from the ontology, e.g. (*scratchy throat*,102618009). Given any pair of tuples $(x_i, y_i), (x_j, y_j) \in \mathcal{X} \times \mathcal{Y}$, the goal of the self-alignment is to learn a function $f(\cdot;\theta) : \mathcal{X} \rightarrow \mathbb{R}^d$ parameterised by θ. Then, the similarity $\langle f(x_i), f(x_j)\rangle$ (in this work we use cosine similarity) can be used to estimate the resemblance of x_i and x_j (i.e., high if x_i, x_j are synonyms and low otherwise). We model f by a BERT model with its output `[CLS]` token regarded as the representation of the input. During the learning process, a sampling procedure selects the informative pairs of training samples and uses them in the pairwise metric learning loss function (introduced shortly).

Online Hard Pairs Mining. The online sample mining finds hard positive/negative pairs or triplets within a mini-batch for efficient training. We use a hard triplet mining condition to filter out "easy positives & negatives" and identify the most informative training triplets. Specifically, we construct all possible triplets for all names within the mini-batch where each triplet is in the form of (x_a, x_p, x_n). Here x_a is called *anchor*, an arbitrary name in the mini-batch; x_p a positive match of x_a (i.e. $y_a = y_p$) and x_n a negative match of x_a (i.e. $y_a \neq y_n$). Among the constructed triplets, we select out all triplets that violate the following condition:

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

where λ is a pre-set margin. In other words, we only consider triplets with the negative sample closer to the positive sample by a margin of λ. These are the hard triplets as their original representations were very far from correct. Every hard triplet contributes one hard positive pair (x_a, x_p) and one hard negative pair (x_a, x_n). We collect all such positive & negative pairs and denote them as $\mathcal{P}, \mathcal{N}$. A similar but not identical triplet mining condition was used by [SKP15] for face recognition in which they only used it for selecting hard negative samples.

Loss Function. We then compute the pairwise cosine similarity of all the names' BERT representations and obtain a similarity matrix $\mathbf{S} \in \mathbb{R}^{|\mathcal{X}_b| \times |\mathcal{X}_b|}$ where each entry $\mathbf{S}_{ij}$ corresponds to the representation cosine similarity between the i-th name in the j-th name in the mini-batch b. We adapted the state-of-the-art metric learning loss of Multi-Similarity loss (MS loss) [WHH+19] for learning from the positive and negative pairs:

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

where α, β are temperature scales; ϵ is an offset applied on the similarity matrix; $\mathcal{P}_i, \mathcal{N}_i$ are indices of positive and negative samples of the *anchor* i.[5] While the first term in Equation (2) pushes negative pairs away from each other while the second term pulls positive pairs together. This dynamic, allows for a re-calibration of the alignment space using the semantic biases of synonymy relations.

3 Experiments

Data and Evaluation Protocol. We experiment across 6 different medical entity linking (MEL) datasets with 4 on the domain of scientific papers: NCBI [DLL14], BC5CDR-c and BC5CDR-d [LSJ+16], MedMentions [ML18], and 2 on the social media domain: COMETA [BLSC20], and AskAPatient [LC16] (see Appendix A.1 for data statistics). We report $\text{Acc}_{@1}$ and $\text{Acc}_{@5}$ (denoted as @1 and @5) for evaluating performance. In all experiments, SapBERT denotes further pre-training with our self-alignment method on UMLS. At the test phase, for all SapBERT models we use nearest neighbour search without further fine-tuning on task data (unless stated otherwise). All reported SapBERT model results are the average of five runs with different random seeds.

Impact of SapBERT. We illustrate the impact of further pre-training with SapBERT on several BERT-based models including the recently proposed PubMedBERT [GTC+20], which is the current SOTA model in the biomedical domain. As is shown in Tab. 1, consistent improvements are obtained across all datasets regardless of the underlying base models.

SapBERT vs. SOTA. We take the best SapBERT setting from Tab. 1 and compare it against various published SOTA results. All SOTA models are trained under supervised learning paradigm, using

[5] We have experimented with other popular loss functions such as InfoNCE [OLV18] and NCA loss [GHRS05] but found our choice is empirically better.

Table 1: Comparison of several BERT-based models w/ or w/o self-alignment pre-training on UMLS. Note that all numbers reported in this table are unsupervised results (not fine-tuned on task data). After pre-training, all models (*BERT+SAPBERT) achieve better performance with statistical significance across all datasets on all metrics. The degree of green indicates the improvement comparing to the base model (the deeper the more).

model	NCBI		BC5CDR-d		BC5CDR-c		MedMentions		AskAPatient		COMETA	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
vanilla BERT [DCLT19]	67.6	77.0	81.4	89.1	79.8	91.2	39.6	60.2	38.2	43.3	40.4	47.7
+ SAPBERT	91.6	95.2	92.7	95.4	96.1	98.0	52.5	72.6	68.4	87.6	59.5	76.8
BIOBERT v1.1 [LYK+20]	71.3	84.1	79.8	92.3	74.0	90.0	24.2	38.5	41.4	51.5	35.9	46.1
+ SAPBERT	91.0	94.7	93.3	95.5	96.6	97.6	53.0	73.7	72.4	89.1	63.3	77.0
BLUEBERT [PYL19]	75.7	87.2	83.2	91.0	87.7	94.1	41.6	61.9	41.5	48.5	42.9	52.9
+ SAPBERT	90.9	94.0	93.4	96.0	96.7	98.2	49.6	73.1	72.4	89.4	66.0	78.8
PUBMEDBERT [GTC+20]	77.8	86.9	89.0	93.8	93.0	94.6	43.9	64.7	42.5	49.6	46.8	53.2
+ SAPBERT	92.0	95.6	93.5	96.0	96.5	98.2	50.8	74.4	70.5	88.9	65.9	77.9

Table 2: SAPBERT vs. SOTA across different MEL datasets, including both scientific papers and social media domain. "-" denotes results not reported in the SOTA paper. "OOM" means out-of-memory (exceeded our 192GB RAM). Blue and red denote unsupervised and supervised models, respectively. **Bold** numbers are the highest in their columns and † indicates statistically significant improvements (T-test, p-value < 0.05) comparing with the supervised SOTA and PUBMEDBERT.

model	scientific language								social media language			
	NCBI		BC5CDR-d		BC5CDR-c		MedMentions		AskAPatient		COMETA	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
supervised SOTA	91.1	93.9	93.2	**96.0**	**96.6**	97.2	OOM	OOM	87.5	-	71.3	77.8
PUBMEDBERT	77.8	86.9	89.0	93.8	93.0	94.6	43.9	64.7	42.5	49.6	46.8	53.2
+ SAPBERT	92.0	**95.6**†	**93.5**†	**96.0**	96.5	**98.2**†	**50.8**†	**74.4**†	70.5	88.9	65.9	77.9
+ FINE-TUNED	**92.3**†	95.5	93.2	95.4	96.5	97.9	50.4	73.9	**89.0**†	**96.2**†	**75.6**†	**84.4**†

task-specific supervision. For scientific language domain and COMETA, the supervised SOTA results are based on BIOSYN [SJLK20], while for AskAPatient dataset the results are based on GEN-RANK [XZB20]. We use red background in Tab. 2 to denote the usage of task-specific labels (i.e., supervised learning or fine-tuning), and blue background to denote fully unsupervised training (e.g., PUBMEDBERT+SAPBERT).

As shown in Tab. 2, measured by Acc$_{@1}$, SAPBERT achieves new SOTA with statistical significance on 5 of the 6 datasets and for the dataset (BC5CDR-c) where SAPBERT is not significantly better, it performs on par with SOTA (96.5 vs. 96.6). Interestingly, on scientific language datasets, SAPBERT outperforms SOTA without any dataset-specific supervision (fine-tuning mostly leads to overfitting and performance drops). On social media language datasets, the unsupervised SAPBERT lags behind supervised SOTA by large margins, highlighting the well-documented complex nature of social media languages [BLSC20, BCL+13]. However, after fine-tuning on the social media MEL datasets (using the same metric learning loss function introduced earlier), SAPBERT outperforms SOTA significantly, indicating that knowledge acquired during the self-aligning pre-training can be adapted to a shifted domain without much effort.

4 Conclusion

In this paper, we presented SAPBERT, a self-alignment pre-training scheme of biomedical entity representations. On the task of medical entity linking, compared with various BERT-based models, we highlighted that the further performance boost achieved by our pre-training method is consistent and independent from the underlying BERT-base used. Our SAPBERT achieves new SOTA in five out of the six datasets. In particular, without even training on task-specific labelled data, SAPBERT outperforms the previous supervised SOTA (sophisticated hybrid entity linking systems) in the scientific language domain. In the future, we plan to (1) generalise such technique to the general domain by leveraging knowledge graphs such as DBpedia; (2) test our model on other BioNLP benchmarks [GTC+20]; (3) extend the current phrase-level representation to sentence-level; and (4) incorporate other entity relations (i.e., hypernymy and hyponymy) into our self-alignment step.

References

[BCL$^+$13] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, 2013.

[BLSC20] Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 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)*, pages 3122–3137, Online, November 2020. Association for Computational Linguistics.

[Bod04] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270, 2004.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[DGJ$^+$19] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McMorran, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2019. *Nucleic Acids Research*, 47:D948–D954, 2019.

[DLL14] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014.

[DN15] Jennifer D'Souza and Vincent Ng. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 297–302, 2015.

[Don06] Kevin Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279, 2006.

[DWRM12] Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012.

[GHRS05] Jacob Goldberger, Geoffrey E Hinton, Sam T Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems*, pages 513–520, 2005.

[GSA$^+$20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. 2020.

[GTC$^+$20] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *arXiv:2007.15779*, 2020.

[HFW$^+$20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[JWX20] Zongcheng Ji, Qiang Wei, and Hua Xu. Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020:269, 2020.

[KdMdY+20] Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. In *International Conference on Learning Representations*, 2020.

[KGH18] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[LC16] Nut Limsopatham and Nigel Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1014–1023, 2016.

[LH18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[LKL+16] Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11:e0164680, 2016.

[LL16] Robert Leaman and Zhiyong Lu. Taggerone: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32:2839–2846, 2016.

[LLG+20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[LSJ+16] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

[LYK+20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[MH08] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[ML18] Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with umls concepts. In *Automated Knowledge Base Construction*, 2018.

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[PST19] Minh C Phan, Aixin Sun, and Yi Tay. Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285, 2019.

[PYL19] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing*, pages 58–65, 2019.

[RSVH15] Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. Overview of the trec 2015 clinical decision support track. In *TREC*, 2015.

[SJLK20] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[SWH15] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.

[TMNM18] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84:93–102, 2018.

[WHH$^+$19] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[WLA$^+$18] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*, 87:12–20, 2018.

[Wri19] Dustin Wright. *NormCo: Deep disease normalization for biomedical knowledge base construction*. PhD thesis, 2019.

[XZB20] Dongfang Xu, Zeyu Zhang, and Steven Bethard. A generate-and-rank framework with semantic type regularization for biomedical concept normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8452–8464, 2020.

A Appendices

A.1 Dataset details

We divide our experimental datasets into two categories (1) scientific language datasests (Appendix A.1.1) where the data is extracted from scientific papers and (2) social media language datasets (Appendix A.1.2) where the data is coming from social media forums like Reddit.

A.1.1 Scientific language datasets

NCBI disease [DLL14] is a corpus containing 793 fully annotated PubMed abstracts and 6,881 mentions. The mentions are mapped into the MEDIC dictionary [DWRM12]. We denote this dataset as "NCBI" in our experiments.

BC5CDR [LSJ$^+$16] consists of 1,500 PubMed articles with 4,409 annotated chemicals, 5,818 diseases and 3,116 chemical-disease interactions. The disease mentions are mapped into the MEDIC dictionary like the NCBI disease corpus. The chemical mentions are mapped into the Comparative Toxicogenomics Database (CTD) [DGJ$^+$19] chemical dictionary. We denote the disease and chemical mention sets as "BC5CDR-d" and "BC5CDR-c" respectively. For NCBI and BC5CDR we use the same data and evaluation protocol by [SJLK20].[6]

MedMentions [ML18] is a very-large-scale entity linking dataset containing over 4,000 abstracts and over 350,000 mentions linked to UMLS 2017AA. According to [ML18], training TaggerOne [LL16], a very popular MEL system, on a subset of MedMentions require >900 GB of RAM. Its massive number of mentions and more importantly the used reference ontology (UMLS 2017AA has over 3 million concepts) make the application of most MEL systems infeasible. However, through our metric learning formulation, SapBert can be applied on MedMentions with minimal effort.

[6] https://github.com/dmis-lab/BioSyn

Table 3: A list of baselines on the 6 different MEL datasets, including both scientific and social media language ones. The last row collects reported numbers from the best performing models. “$*$” denotes results produced using official released code. “-” denotes results not reported in the cited paper. “OOM” means out-of-memoery.

model	NCBI		BC5CDR-d		BC5CDR-c		MedMentions		AskAPatient		COMETA	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
SIEVE-BASED [DN15]	84.7	-	84.1	-	90.7	-	-	-				
WORDCNN [LC16]	-	-	-	-	-	-	-	-	81.4	-	-	-
WORDGRU+TF-IDF [TMNM18]	-	-	-	-	-	-	-	-	85.7	-	-	-
TAGGERONE [LL16]	87.7	-	88.9	-	94.1	-	OOM	OOM	-	-	-	-
NORMCO [Wri19]	87.8	-	88.0	-	-	-	-	-	-	-	-	-
BNE [PST19]	87.7	-	90.6	-	95.8	-	-	-	-	-	-	-
BERTRANK [JWX20]	89.1	-	-	-	-	-	-	-	-	-	-	-
GEN-RANK [XZB20]	-	-	-	-	-	-	-	-	**87.5**	-	-	-
BIOSYN [SJLK20]	**91.1**	**93.9**	**93.2**	**96.0**	**96.6**	**97.2**	OOM	OOM	82.6*	87.0*	71.3*	77.8*
DICT+SOILOS+NEURAL [XZB20]	-	-	-	-	-	-	-	-	-	-	**79.0**	-
supervised SOTA	91.1	93.9	93.2	96.0	96.6	97.2	OOM	OOM	87.5	-	79.0	-

A.1.2 Social media language datasets

COMETA [BLSC20] is a recently released large-scale MEL dataset that specifically focuses on MEL in the social media domain, containing around 20k medical mentions extracted from health-related discussions on `reddit.com`. Mentions are mapped to SNOMED CT [Don06]. We use the “stratified (general)” split and follow the evaluation protocol of the original paper.[7]

AskAPatient [LC16] includes 17,324 adverse drug reaction (ADR) annotations collected from `askapatient.com` blog posts. The mentions are mapped to 1,036 medical concepts grounded onto SNOMED CT. For this dataset, we follow the 10-fold evaluation protocol stated in the original paper.[8]

A.2 Data preparation details for UMLS pre-training

We download the UMLS 2020 AA version from `https://download.nlm.nih.gov/umls/kss/2020AA/umls-2020AA-full.zip`. We then extract all English entries from the raw files and convert all entity names into lowercase (duplications are removed). Besides all the officially defined ones, we also include tradenames of drugs as synonyms.

A.3 Training details

Pre-training. To make sure that sufficient positive pairs exist in every mini-batch, we batch pairs of positive names instead of batching completely random individual samples. Specifically, we generate a positive pair list offline using the (name, CUI) list obtained in Appendix A.2. We enumerate all the possible combinations of any two names sharing the same CUI and all such pairs constitute the positive pair list. For balanced training, any concepts with more than 50 positive pairs are randomly trimmed to 50 pairs. During training, we use AdamW [LH18] with a learning rate of `2e-5` and weight decay rate of `1e-2`. Models are trained on the prepared UMLS data for 1 epoch with a batch size of 480 (approximately 50k iterations). This takes approximately 5 hours on 2 GTX 2080Ti.

Fine-tuning. Similar to pre-training, a positive pair list is generated through traversing the combinations of mention and all ground truth synonyms where mentions are from the training set and ground truth synonyms are from the reference ontology. We use the same optimiser and learning rates but train with a batch size of 256 (to accommodate memory of 1 GPU). On scientific language datasets, we train for only 1 epoch while on social media ones we train for 15 epochs.

A.4 Table of supervised baseline models.

The full table of supervised baseline models is provided in Tab. 3.

[7] `https://www.siphs.org/corpus`

[8] `https://zenodo.org/record/55013`