Experiment Protocol

Will Thompson

October 19, 2022

1 Introduction

Entity linking is the task of assigning unique entity identifiers to mentions of entities in text [Sev+22]. There are two core data sources for this task: (1) a collection of unstructured documents $\mathcal{D}$ that contain a set $\mathcal{M}$ of text spans interpreted as entity mentions, and (2) a structured domain knowledge base $\mathcal{K}$ that contains a set $\mathcal{E}$ of uniquely identified entities, which can be anything of interest. Entity linking is the task of defining a function $\Gamma : \mathcal{M}X\mathcal{D} \rightarrow \mathcal{E}$ that maps a mention $m_i \in \mathcal{M}$ in document $d_j \in \mathcal{D}$ to a unique entity $e_k \in \mathcal{E}$.

My core research interest is in applying entity linking to *clinical notes*, unstructured text that can be found in large quantities in electronic health records (EHR). There is a lot of potentially useful information in such notes, ranging from disease mentions, descriptions of symptoms, past medical history, response to treatments, and so on. While EHR also have large quantities of structured information, it turns out that the structured information that is available often provides an incomplete (or even misleading) picture of a patient's clinical status. Notes in the EHR typically contain a more direct assessment of a patient's condition, and can be the sole source of information for areas such as medical symptoms, treatment response, cancer grading, and past medical history. There is a huge amount of information locked up in clinical notes, and clinical entity linking has many potentially use cases for research into diseases, measuring population health, creation of disease-specific registries, and clinical trial recruitment.

My general NLP-related research goal is to improve on the state of the art for clinical entity linking – linking mentions of diseases, symptoms, treatments (etc.) in clinical notes to entity identifiers in domain knowledge bases such as the Unified Medical Language System (UMLS; [Bod04]), a large domain knowledge base created by the National Library of Medicine containing over 4 million clinically relevant entities linked into a semantic network. There are two highly salient background facts relevant to clinical entity linking. First, there is a general lack of labeled data. Due to privacy concerns and regulations (such as HIPAA), there are few publicly available corpora, and even fewer that are annotated with gold standard labels. Second, there is an abundance of knowledge bases such as the UMLS.

These two facts lead me to focus on *self supervised* [KRT22] approaches to entity linking. These are approaches that leverage domain knowledge bases (such as the UMLS) to automatically generate candidate entity mentions to train entity linking models. Self-supervision has been successfuly used to overcome the general lack of annotated resources for the clinical domain. In particular for this paper, I will focus on the recent model `KRISSBERT`, proposed by [Zha+22]. This model has achieved state of the art for self-supervised approaches to biomedical entity linking. This will be the reference approach and baseline model for the research ideas proposed in this paper.

2 Hypotheses

My hypotheses are stated relative to a self-supervised approach to entity linking, with the following core assumptions:

- **Assumption 1**: There are two data sources consisting of a collection of unlabeled text and a list of structured entities, which each have a unique identifer and a canonical name.
- **Assumption 2**: Self-supervised learning uses the entity names to generate entity mentions for training.

  For the purposes of this paper, I will make the following additional assumptions:

- **Assumption 3**: There are multiple target domains $\mathcal{K}^{(i)}$, subsets of $\mathcal{K}$. Each target domain $\mathcal{K}^{(i)}$ contains a set of entities $\mathcal{E}^{(i)}$ that are subsets of $\mathcal{E}$. For example, if we use the UMLS as our knowledge base, there are subsets that describe cancer diseases, or heart disease, etc. Each of these target domains may also be affiliated with a specific collection of documents $\mathcal{D}^{(j)}$ subset of $\mathcal{D}$. For example, we may be interested in extracting cancer concepts from pathology notes, which have a much different structure and content than patient discharge notes.
- **Assumption 4**: The domain $\mathcal{E}^{(i)}$ may or may not have an annotated gold standard available for model training and testing. Where available, they can potentially be used to improve results derived solely from self-supervised models.

Given these assumptions, the following hypotheses relate to ways in which self-supervised entity linking models can be augmented and potentially improved:

- **Hypothesis 1**: Relative to Assumptions (1-2), entity linking accuracy can be increased by generating higher-quality entity mention examples for training. [Zha+22] use exact string matching from texts to canonical

entity names in the UMLS to generate training examples. They spot-checked a random sample and determined that this method achieved approximately 85% accuracy (given the context, I'm assuming this means 0.85 *precision* for linked entities from the generated mentions). Although this is quite good, and given the sheer volume of data they generate it performs well in practice, my hypothesis is we could do better by placing further restrictions on string matching – such as requiring a minimum token length to reduce potential linguistic ambiguities. Reducing training noise will (hopefully) result in better model performance.

- **Hypothesis 2**: Relative to Assumptions (1-3), entity linking accuracy can be increased by using a domain-specific target corpus for pre-training or fine-tuning. [Zha+22] use PubMedBERT [Gu+22], trained on large amounts of PubMed abstracts. This is naturally a good corpus to use for the target domain of biomedical literature, but it might not be the best data source for clinical notes, which are drastically different in terms of content and structure. There are pre-existing language models trained on clinical notes (such as ClinicalBERT [HAR20]) which may serve as a better foundation, and we can experiment with using very specific subsets of notes (such as pathology or radiology notes) to fine-tune to specific target domains.

- **Hypothesis 3**: At inference time, it appears to be common practice (e.g., [Zha+22], [LJ20], [Log+19], [Wu+20]) to use a nearest neighbour approach to select the top-K entities that are potentially best matches to a given entity mention. Relative to Assumptions (1-3), entity linking accuracy can be improved for specific target domains by pre-pruning the entity space for target-specific concepts, before performing nearest neighbour search. This seems like it couldn't possibly hurt (assuming we only care about the target concepts for a particular use case), and would very likely improve entity linking performance by removing irrelevant distractors.

- **Hypothesis 4**: Relative to Assumptions (1-4), entity linking accuracy can be increased by using labeled examples for a target domain to fine-tune the self-supervised model. This again seems like it couldn't possibly hurt, and the question is to what degree it might improve accuracy.

3 Datasets

I will use the following datasets, each of which I have already accessed:

1. UMLS [Bod04]: a large domain knowledge base created by the National Library of Medicine containing over 4 million clinically relevant entities linked into a semantic network.

2. MedMentions [ML19]: the largest available dataset mapping text to UMLS concepts. The texts consist of over 4,000 PubMed abstracts containing

over 350,000 linked mentions. This dataset is widely used in the literature on biomedical entity linking.

3. MIMIC-III [Gol+00]: a large database of deidentified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database includes a clinical notes table with over 2 million rows.

4. Shared Annotated Resources (ShARe) SemEval 2015 corpus [Sav], [Pra+14]: 531 deidentified clinical discharge summaries and radiology reports from the MIMIC clinical database. The ShARe corpus contains gold-standard annotations of disorder mentions and a set of attributes.

5. RadGraph [Jai+]: a dataset of entities and relations in full-text radiology reports extracted from MIMIC chest x-rays. RadGraph contains a test set of 500 annotated notes, and a test set of 100 notes. Mention types fall under the two broad categories of Observation and Anatomy.

4 Metrics

I will use two quantitative metrics:

1. Top-K accuracy: this metric computes the number of times where the correct entity label is among the top k labels predicted. Top-1 accuracy will be my primary metric for evaluating systems.

2. Mean reciprocal rank (MRR): MRR measures how far down the ranking the correct entity is. This will give a sense of "how far off" (on average) the true results are from being correctly predicted.

In addition to using these quantitative metrics, I will be doing a qualitative error analysis to get a better sense of what is going on with the entity linking systems. I believe it will be particularly instructive to look at entities with ambiguous entity mentions (names that are shared across entity types). I will also be doing an ablation analysis to look at the specific contributions of the various components.

5 Models

My baseline model will be `KRISSBERT` [Zha+22] (see Figure 1). This self-supervised model is trained using the entire UMLS domain model. They compile a complete list of UMLS entity names into a trie, and use it to efficiently search text for entity mentions; they applied this string matching approach to PubMed abstracts and used it to generate 1.6 billion mention examples, each of which is uniquely linked to a UMLS entity concept identifier. They then use these mentions to train an entity encoder using contrastive loss [OLV19], initialized with PubMedBERT [Gu+22]. They call the

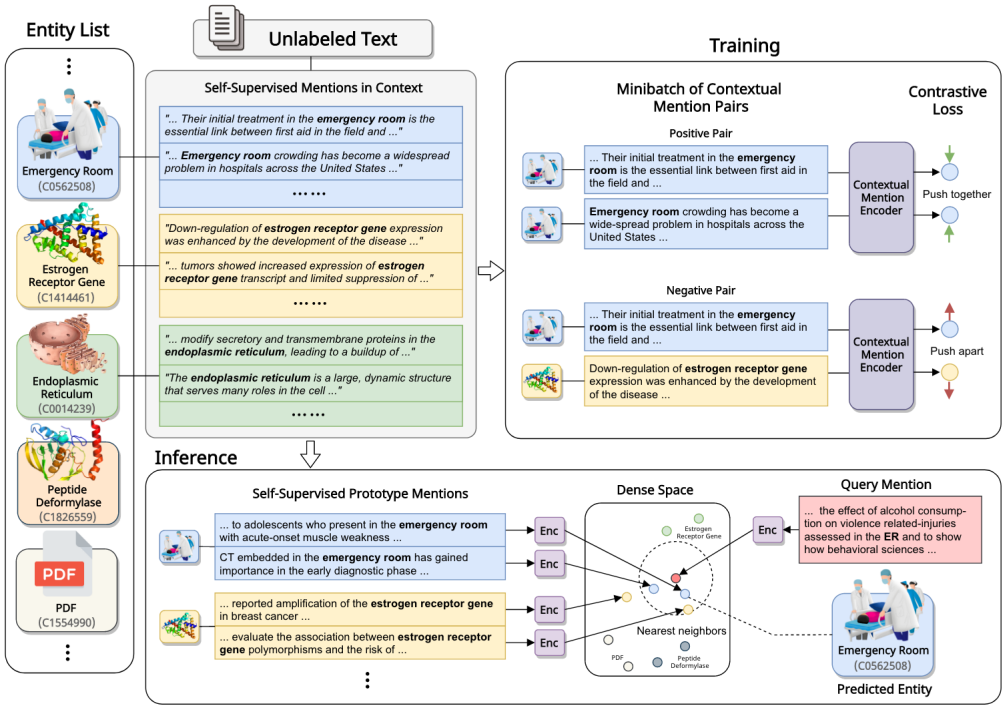

Figure 1: From [Zha+22]

end result KRISSBERT, which is available for download on the Huggingface Hub (microsoft/BiomedNLP-KRISSBERT-PubMed-UMLS-EL). Inference is performed using a efficient and scalable form of nearest neighbor search [JDJ17]. They sample self-supervised mentions as prototypes for each entity and map the test mention to the most similar prototype.

Relation to datasets: The baseline KRISSBERT model is specifically designed to handle biomedical and clinical entity linking scenarios. My experiments will tie together this model with the datasets described above as follows:

- I will first use PubMed and UMLS to re-implement the KRISS training algorithm and validate it against the MedMentions corpus, to see if I can reproduce the results for this dataset in [Zha+22].
- I will then use the UMLS and the MIMIC-III notes table to generate a more clinically specific version of the model.
- I will use the ShARe and RadGraph datasets to test Hypotheses (3) and (4).

In relationship to the metrics described in the Metrics section, these will be used to measure performance on every version of clinical entity linking model that is generated to test the hypotheses.

6 General reasoning

The general approach is to augment self-supervised entity linking models with target specific domain information that can be used to improve entity linking performance, as measured by top-K accuracy and mean reciprocal rank. The baseline model is KRISS, which exhibits state of the art performance for self-supervised entity linking in the biomedical domain. Applying this approach to clinical notes, I believe we can do better by using clinically-focused resources, such as using ClinicalBERT [HAR20] as the pre-trained model for mention embeddings, and filtering entities to smaller subsets of the UMLS for targeted use cases. Targeted datasets (such as radiology notes annotated with chest x-ray concepts) can be used to both fine-tune the mention embeddings model and improve inference by focusing just on concepts relevant to the use case.

7 Summary of progress so far

Completed steps:

1. ✓Obtained access to the UMLS and have installed it on a local postgresql database on my macbook laptop.
2. ✓Obtained the MedMentions dataset and written python code to parse it and convert it into a Pandas dataframe for ease of manipulation.

3. ✓Obtained access to the MIMIC 3 note dataset, which required proper human subjects training. Physionet (the organization that distributes this dataset) also makes available a variety of other corpora, including RadGraph and ShaREe, which I have full access to through my Physionet account.

4. ✓The `KRISSBERT` model and some test code are available through Huggingface (`huggingface.co/microsoft/BiomedNLP-KRISSBERT-PubMed-UMLS-EL`). I've cloned this code and executed it to generate prototype embeddings and run entity linking against the MedMentions corpus. Unfortunately, it does not appear that the authors have released the model training code for this project.

Remaining steps:

1. Implement the training algorithm as described in the paper

2. Modify the entity mention generation algorithm to generate high-quality entity mentions, by imposing token length restrictions (and potentially other restrictions, based on analysis of generation errors).

3. Evaluate performance of entity linker on target clinical domains (RadGraph, ShARe), using techniques described above for using a different base model, fine-tuning, pruning the entity set, etc.

Concerns:

1. My biggest concern is the ability to implement the `KRISS` training algorithm faithfully, based just on the details in the paper. I suspect this will be challenging and time-consuming.

2. An easier task to envision is the entity pruning for targeted inference. Just implementing this would not require any changes to the `KRISSBERT` model, which I could simply use as is.

3. Perhaps the easiest task (thanks to Huggingface) is fine-tuning `KRISSBERT` with target specfic data, such as all of MIMIC-III notes, or just the radiology notes, etc.

References

[Gol+00] Ary L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals". In: *Circulation* 101.23 (June 13, 2000). ISSN: 0009-7322, 1524-4539. DOI: 10.1161/01.CIR.101.23.e215. URL: https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215 (visited on 10/06/2022).

[Bod04] O. Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology". In: *Nucleic Acids Research* 32.90001 (Jan. 1, 2004), pp. 267D–270. ISSN: 1362-4962. DOI: 10.1093/nar/gkh061. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh061 (visited on 10/05/2022).

[Pra+14] Sameer Pradhan et al. "SemEval-2014 Task 7: Analysis of Clinical Text". In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014, pp. 54–62. DOI: 10.3115/v1/S14-2007. URL: http://aclweb.org/anthology/S14-2007 (visited on 10/19/2022).

[JDJ17] Jeff Johnson, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs". In: (2017). Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1702.08734. URL: https://arxiv.org/abs/1702.08734 (visited on 10/19/2022).

[Log+19] Lajanugen Logeswaran et al. *Zero-Shot Entity Linking by Reading Entity Descriptions*. June 17, 2019. arXiv: 1906.07348[cs]. URL: http://arxiv.org/abs/1906.07348 (visited on 10/04/2022).

[ML19] Sunil Mohan and Donghui Li. *MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts*. Feb. 25, 2019. arXiv: 1902.09476[cs]. URL: http://arxiv.org/abs/1902.09476 (visited on 10/19/2022).

[OLV19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*. Jan. 22, 2019. arXiv: 1807.03748[cs,stat]. URL: http://arxiv.org/abs/1807.03748 (visited on 10/19/2022).

[HAR20] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. Nov. 28, 2020. arXiv: 1904.05342[cs]. URL: http://arxiv.org/abs/1904.05342 (visited on 10/04/2022).

[LJ20] Daniel Loureiro and Alípio Mário Jorge. "MedLinker: Medical Entity Linking with Neural Representations and Dictionary Matching". In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Vol. 12036. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 230–237. ISBN: 978-3-030-45441-8 978-3-030-45442-5. DOI: 10.1007/978-3-030-45442-5_29. URL: http://link.springer.com/10.1007/978-3-030-45442-5_29 (visited on 10/04/2022).

[Wu+20] Ledell Wu et al. *Scalable Zero-shot Entity Linking with Dense Entity Retrieval*. Sept. 29, 2020. arXiv: 1911.03814[cs]. URL: http://arxiv.org/abs/1911.03814 (visited on 10/06/2022).

[Gu+22] Yu Gu et al. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing". In: *ACM Transactions on Computing for Healthcare* 3.1 (Jan. 31, 2022), pp. 1–23. ISSN: 2691-1957, 2637-8051. DOI: 10.1145/3458754. arXiv: 2007.15779[cs]. URL: http://arxiv.org/abs/2007.15779 (visited on 10/19/2022).

[KRT22] Rayan Krishnan, Pranav Rajpurkar, and Eric J. Topol. "Self-supervised learning in medicine and healthcare". In: *Nature Biomedical Engineering* (Aug. 11, 2022). ISSN: 2157-846X. DOI: 10.1038/s41551-022-00914-1. URL: https://www.nature.com/articles/s41551-022-00914-1 (visited on 10/18/2022).

[Sev+22] Ozge Sevgili et al. "Neural Entity Linking: A Survey of Models Based on Deep Learning". In: *Semantic Web* 13.3 (Apr. 6, 2022), pp. 527–570. ISSN: 22104968, 15700844. DOI: 10.3233/SW-222986. arXiv: 2006.00575[cs]. URL: http://arxiv.org/abs/2006.00575 (visited on 10/04/2022).

[Zha+22] Sheng Zhang et al. *Knowledge-Rich Self-Supervision for Biomedical Entity Linking*. May 23, 2022. arXiv: 2112.07887[cs]. URL: http://arxiv.org/abs/2112.07887 (visited on 10/17/2022).

[Jai+] Saahil Jain et al. *RadGraph: Extracting Clinical Entities and Relations from Radiology Reports*. Version Number: 1.0.0 Type: dataset. DOI: 10.13026/HM87-5P47. URL: https://physionet.org/content/radgraph/1.0.0/ (visited on 10/19/2022).

[Sav] Guergana Savova. *Analysis of Clinical Text: Task 14 of SemEval 2015*. Version Number: 2.0 Type: dataset. DOI: 10.13026/61RG-Q298. URL: https://physionet.org/content/semeval2015/2.0/ (visited on 10/19/2022).