Knowledge-Rich Self-Supervision for Biomedical Entity Linking

Sheng Zhang[*] **Hao Cheng**[*] **Shikhar Vashishth**[*] **Cliff Wong** **Jinfeng Xiao**[†]
Xiaodong Liu **Tristan Naumann** **Jianfeng Gao** **Hoifung Poon**
Microsoft Research
[†]University of Illinois at Urbana-Champaign

Abstract

Entity linking faces significant challenges such as prolific variations and prevalent ambiguities, especially in high-value domains with myriad entities. Standard classification approaches suffer from the annotation bottleneck and cannot effectively handle unseen entities. Zero-shot entity linking has emerged as a promising direction for generalizing to new entities, but it still requires example gold entity mentions during training and canonical descriptions for all entities, both of which are rarely available outside of Wikipedia. In this paper, we explore Knowledge-RIch Self-Supervision (`KRISS`) for biomedical entity linking, by leveraging readily available domain knowledge. In training, it generates self-supervised mention examples on unlabeled text using a domain ontology and trains a contextual encoder using contrastive learning. For inference, it samples self-supervised mentions as prototypes for each entity and conducts linking by mapping the test mention to the most similar prototype. Our approach can easily incorporate entity descriptions and gold mention labels if available. We conducted extensive experiments on seven standard datasets spanning biomedical literature and clinical notes. Without using any labeled information, our method produces `KRISSBERT`, a universal entity linker for over three million UMLS entities that attains new state of the art, outperforming prior self-supervised methods by as much as 20 absolute points in accuracy. We released `KRISSBERT` at `https://aka.ms/krissbert`.

1 Introduction

Entity linking maps mentions to unique entities in a target knowledge base (Roth et al., 2014). It can be viewed as the extreme case of named entity recognition and entity typing, where the category number swells to tens of thousands or even millions. Entity linking is particularly challenging in high-value domains such as biomedicine, where variations and ambiguities abound. For instance, depending on the context, *"PDF"* may refer to a gene (Peptide Deformylase, Mitochondrial), or file type (Portable Document Format). Similarly, *"ER"* could refer to emergency room, the organelle endoplasmicreticulum, or the estrogen receptor gene. Moreover, the number of entities in domains such as biomedicine can be very large. The Unified Medical Language System (UMLS), a representative ontology for biomedicine, contains over three million entities (Bodenreider, 2004).

Standard classification approaches such as MedLinker (Loureiro and Jorge, 2020) require example gold mentions for each entity and cannot effectively handle new entities for which there are no labeled examples in training. Recently, zero-shot entity linking has emerged as a promising direction for generalizing to unseen entities (Logeswaran et al., 2019; Wu et al., 2020), by learning to encode contextual mentions for similarity comparison against reference entity descriptions. Existing methods, however, require example gold entity mentions during training, as well as canonical descriptions for all entities. While applicable to Wikipedia entities, these methods are hard to generalize to other domains, where such labeled information is rarely available at scale.

In this paper, we explore **K**nowledge-**RI**ch **S**elf-**S**upervision (`KRISS`) for entity linking by leveraging readily available domain knowledge to compensate for the lack of labeled information (Figure 1). For entity linking, the most relevant knowledge source is the domain ontology. The core of an ontology is the entity list, which specifies the unique identifier and canonical name for each entity and is the prerequisite for entity linking. Our method only requires the entity list and unlabeled text, which are readily available in any domain.

In training, `KRISS` uses the entity list to generate self-supervised mention examples from unlabeled

[*] These authors contributed equally to this research.
[†] Work done as an intern at Microsoft Research.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 868 - 880
December 7-11, 2022 ©2022 Association for Computational Linguistics

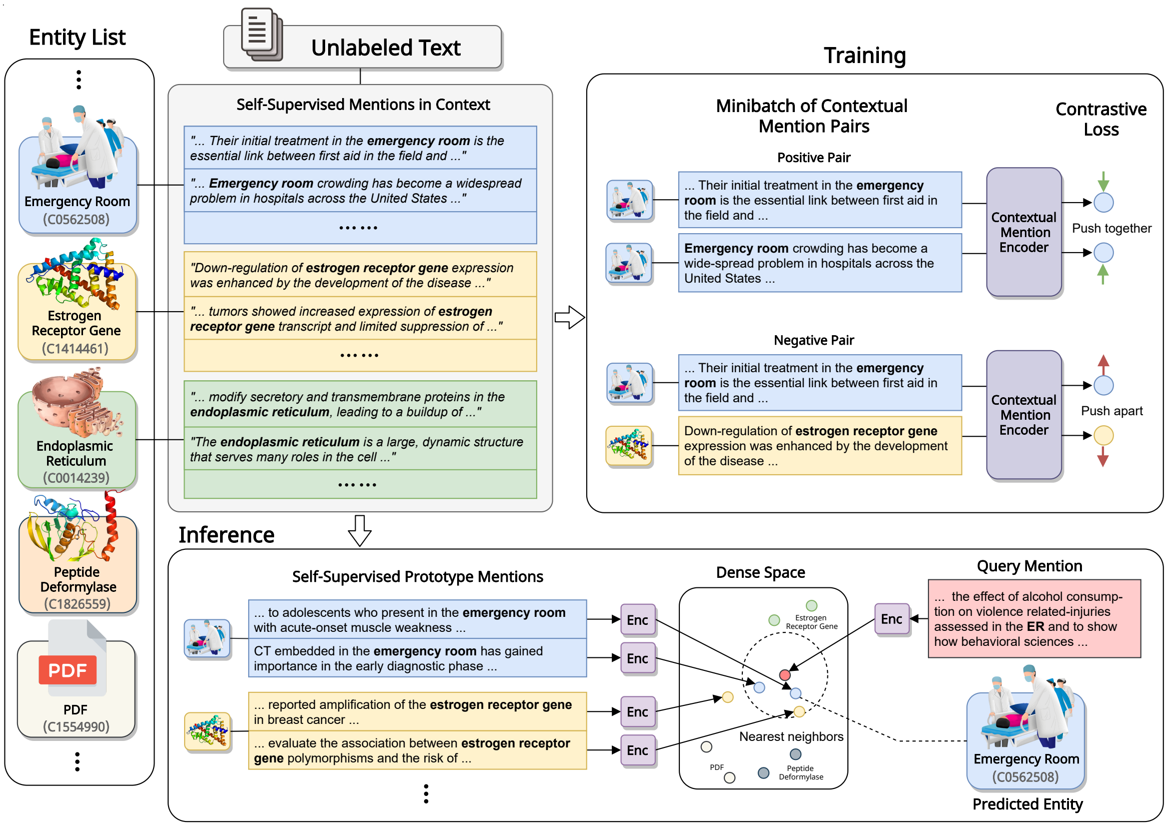

Entity List
Emergency Room
(C0562508)
Estrogen Receptor Gene
(C1414461)
Endoplasmic Reticulum
(C0014239)
Peptide Deformylase
(C1826559)
PDF
PDF
(C1554990)
Unlabeled Text
Self-Supervised Mentions in Context
"... Their initial treatment in the emergency room is the essential link between first aid in the field and ..."
"... Emergency room crowding has become a widespread problem in hospitals across the United States ..."
"Down-regulation of estrogen receptor gene expression was enhanced by the development of the disease ..."
"... tumors showed increased expression of estrogen receptor gene transcript and limited suppression of ..."
"... modify secretory and transmembrane proteins in the endoplasmic reticulum, leading to a buildup of ..."
"The endoplasmic reticulum is a large, dynamic structure that serves many roles in the cell ..."
Training
Minibatch of Contextual Mention Pairs
Contrastive Loss
Positive Pair
... Their initial treatment in the emergency room is the essential link between first aid in the field and ...
Emergency room crowding has become a wide-spread problem in hospitals across the United States ...
Contextual Mention Encoder
Push together
Negative Pair
... Their initial treatment in the emergency room is the essential link between first aid in the field and ...
Down-regulation of estrogen receptor gene expression was enhanced by the development of the disease ...
Contextual Mention Encoder
Push apart
Inference
Self-Supervised Prototype Mentions
... to adolescents who present in the emergency room with acute-onset muscle weakness ...
CT embedded in the emergency room has gained importance in the early diagnostic phase ...
... reported amplification of the estrogen receptor gene in breast cancer ...
... evaluate the association between estrogen receptor gene polymorphisms and the risk of ...
Enc
Enc
Enc
Enc
Dense Space
Estrogen Receptor Gene
Nearest neighbors
PDF
Peptide Deformylase
Query Mention
Enc
... the effect of alcohol consumption on violence related-injuries assessed in the ER and to show how behavioral sciences ...
Emergency Room
(C0562508)
Predicted Entity

Figure 1: Illustration of knowledge-rich self-supervised entity linking.

text, and trains a contextual mention encoder using contrastive learning (Gao et al., 2014; Wu et al., 2020), by mapping mentions of the same entity closer. For inference, KRISS samples prototypes for each entity from the self-supervised mentions. Given a test mention, KRISS finds the most similar prototype and returns the entity it represents.

Prior methods that leverage domain ontology for entity linking often resort to string matching (against entity names and aliases), making them vulnerable to both variations and ambiguities. Recently, a flurry of methods have been proposed to conduct biomedical entity representation learning from synonyms in the ontology, such as BIOSYN (Sung et al., 2020), SapBERT (Liu et al., 2021), and others (Lai et al., 2021). These methods can resolve variations to some extent, but they completely ignore mention contexts and cannot resolve ambiguities. Given an entity mention, they only predict a surface form, rather than a unique entity as required by entity linking (e.g., see footnote 2 in SapBERT (Liu et al., 2021)). As we will show in §4.5, their predicted surface forms are often ambiguous and can't be mapped to a unique entity. Unfortunately, starting from BIOSYN, these papers all adopt an incorrect evaluation method that simply ignores the ambiguity and declares the predicted surface form as correct. Consequently, their reported "entity linking" scores are often highly inflated and do not represent true linking performance. In §4.5, we provide a detailed analysis to illustrate this problem, which we hope would contribute to rectifying this significant evaluation error in future entity linking work.

We conduct our study on biomedicine, which serves as a representative high-value domain where prior methods are hard to apply. Among the three million biomedical entities in UMLS, less than 6% have any description available. Gold mention labels are available for only a tiny fraction of entities. E.g., MedMentions (Mohan and Li, 2019), the largest biomedical entity linking dataset, only covers 35 thousand entities.

We applied our method to train KRISSBERT, a universal entity linker for all three million biomedical entities in UMLS, using only the entity list in UMLS and unlabeled text in PubMed[1]. KRISSBERT can also incorporate additional domain knowledge in UMLS such as entity aliases and ISA hierarchy. We conducted extensive evaluation on

[1] https://pubmed.ncbi.nlm.nih.gov/

seven standard biomedical entity linking datasets spanning biomedical literature and clinical notes. KRISSBERT demonstrated clear superiority, outperforming prior state of the art by 10 points in average accuracy and by over 20 points in MedMentions.

KRISSBERT can be directly applied to lazy learning (§3.7) with no additional training, by simply using gold mention examples as prototypes during inference. This universal model already attains comparable results as dataset-specific state-of-the-art supervised methods, each tailored to an individual dataset by limiting entity candidates and using additional supervision sources and more complex methods (e.g., coreference rules and joint inference). We released KRISSBERT to facilitate research and applications in biomedical entity linking.

2 Related Work

Entity linking Many applications require mapping mentions to unique entities. E.g., knowing that *some drug* can treat *some disease* is not very useful, unless we know the specific drug and disease. Entity linking is inherently challenging given the large number of unique entities. Prior work often adopts a pipeline approach that first narrows entity candidates to a small set (candidate generation) and then learns to classify contexts of the mention and a candidate entity (candidate ranking) (Bunescu and Paşca, 2006; Cucerzan, 2007; Ratinov et al., 2011). Candidate generation often resorts to string matching or TF-IDF variants (e.g., BM25), which are vulnerable to variations. Ranking features are manually engineered or learned via various neural architectures (He et al., 2013; Ganea and Hofmann, 2017; Kolitsas et al., 2018). Additionally, entity relations (e.g., concept hierarchy) and joint inference have been explored for improving accuracy (Gupta et al., 2017; Murty et al., 2018; Cheng and Roth, 2013; Le and Titov, 2018). These methods are predominantly supervised, and suffer from the scarcity of annotated examples, especially given the large number of entities to cover. By contrast, KRISSBERT leverages self-supervision using readily available domain knowledge and unlabeled text, and can effectively resolve variations and ambiguities for millions of entities.

Knowledge-rich self supervision Domain ontology such as UMLS has been applied to self-supervise biomedical named entity recognition (Zhang and Elhadad, 2013; Almgren et al., 2016). Recently, Sung et al. (2020); Liu et al. (2021) propose SapBERT for mention normalization by conducting contrastive learning over synonyms from UMLS. However, SapBERT completely ignores mention contexts. It can resolve some variations but not ambiguity[2]. By contrast, we apply contrastive learning on mention contexts, and leverage unlabeled text to generate self-supervised examples. SapBERT relies on synonyms to learn spelling variations. Our approach can learn with just the canonical name for each entity, as self-supervised mention examples naturally capture contexts where synonymous mentions may appear in.

3 Knowledge-Rich Self-Supervision for Entity Linking

Entity linking grounds textual mentions to unique entities in a given database/dictionary. Formally, the goal of entity linking is to learn a function $\texttt{Link}: (m, T) \rightarrow e$ that maps mention m in the context T to the unique entity e. *Self-supervised entity linking* assumes no access to any gold mention examples. The knowledge-rich self-supervision setting (KRISS) assumes that only a domain ontology $\mathcal{O}$ and an unlabeled text corpus $\mathcal{T}$ are available. In particular, we require the availability of an entity list, which specifies for each entity the unique identifier and a canonical name. Entity list is the prerequisite for entity linking, as it provides the targets for linking. Our framework can also incorporate other knowledge in the ontology (§3.5).

3.1 Generating Self-Supervision

To generate self-supervised mention examples, we first compile a list of entity names from preferred terms in UMLS. We then build a trie from these names (case preserved) to efficiently search them in plain text. When an exact match is found, a fixed-size window around the mention will be returned as context. Some preferred terms are shared by multiple entities. To reduce noise for training and inference, we skip the ambiguous terms. We conducted this process on PubMed abstracts and obtained over 1.6 billion mention examples, each of which is uniquely linked to an entity in UMLS. The estimated linking accuracy based on random samples is 85%. Note that not all UMLS entities have self-supervised examples, as they have never been mentioned in PubMed. This is not an issue

[2] The SapBERT paper states:"*In this work, biomedical entity refers to the surface forms of biomedical concepts*". As aforementioned, many surface forms of biomedical entities are highly ambiguous (e.g., "*PDF*", "*ER*").

for training as our goal is to learn a general encoder that maps mentions of the same entity closer (§3.2). For inference, the `ISA` hierarchy in UMLS can be leveraged to compensate for the lack of self-supervised examples (§3.5).

3.2 Contrastive Learning

Given the self-supervised mentions, we train a mention encoder using contrastive learning by mapping mentions of the same entity closer and mentions of different entities farther apart. Specifically, each mention m is encoded into a contextual vector $\mathbf{c}$ using a Transformer-based encoder (Vaswani et al., 2017), with the following input format:

$$\texttt{[CLS]}\ \text{ctx}_l\ \texttt{[M}_s\texttt{]}\ \text{mention}\ \texttt{[M}_e\texttt{]}\ \text{ctx}_r\ \texttt{[SEP]}$$

where ctx_l and ctx_r denote the left and right context respectively; `[M`$_s$`]` and `[M`$_e$`]` are markers indicating the start and end of the mention; `[CLS]` and `[SEP]` are special encoding tokens. The last-layer hidden state of `[CLS]` is used as the contextual vector $\mathbf{c}$. See subsection A.3 for an illustration.

In a minibatch, we sample $2N$ self-supervised mentions from N entities and encode them into contextual vectors $\{\mathbf{c}_1, ..., \mathbf{c}_{2N}\}$, where $\mathbf{c}_{2k-1}$ and $\mathbf{c}_{2k}$ are from the same entity. Given a positive pair $(\mathbf{c}_i, \mathbf{c}_j)$, we treat the other $2(N-1)$ vectors within a minibatch as negative examples, and compute the InfoNCE loss (Oord et al., 2018) as:

$$\ell_{\mathbf{c}_i,\mathbf{c}_j} = -\log \frac{\exp(\mathbf{c}_i^\top \mathbf{c}_j/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathbf{c}_i^\top \mathbf{c}_k/\tau)}, \quad (1)$$

where $\mathbb{1}_{[k\neq i]} \in \{0,1\}$ is an indicator function evaluating to 1 iff $k \neq i$ and τ denotes a temperature parameter. The final loss is computed across all positive pairs in a minibatch:

$$\mathcal{L} = \frac{1}{2N}\sum_{k=1}^{N}[\ell_{\mathbf{c}_{2k-1},\mathbf{c}_{2k}} + \ell_{\mathbf{c}_{2k},\mathbf{c}_{2k-1}}] \quad (2)$$

3.3 Mention Masking and Replacement

Skipping ambiguous names improves the quality of mention examples (§3.1), but models trained with such self-supervision tend to over-index on surface matching, limiting generalizability. To overcome this, we propose two strategies to augment alternative views of the encoder input during training:

Mention Masking With a probability p_{mask}, we mask the mention using `[MASK]`, which regularizes the model from lexical memorization and encourages it to leverage cues from surrounding context.

Mention Replacement With a probability p_{replace}, the mention is replaced with its synonym in UMLS while the context is kept unchanged. This yields a new mention of the same entity, encouraging the model to generalize across entity variations.

3.4 Linking with Self-Supervised Prototypes

At test time, for each entity e in the entity list $\mathcal{E}$ compiled in §3.1, we sample a small set of self-supervised mentions as reference prototypes, denoted as $\texttt{Proto}(e)$. Given a test/query mention m_q, we return the entity e with the most similar prototype m_p based on the self-supervised encoding:

$$\texttt{Link}(m_q) = \underset{e\in\mathcal{E}}{\text{argmax}} \max_{m_p\in\texttt{Proto}(e)} \mathbf{c}_q^\top \mathbf{c}_p \quad (3)$$

For efficient linking, we pre-compute the contextual vectors of all reference prototypes and leverage fast nearest neighbor search tool that can scale to millions of entities (Johnson et al., 2019).

3.5 Incorporating Additional Knowledge

Our self-supervised entity linking formulation can easily incorporate other knowledge available in an ontology, either by generating additional mention examples from unlabeled text, or by creating special entity-centric examples, which can be used both for learning and inference. This is especially important for entities without self-supervised mentions from PubMed (§3.1).

Aliases Ontology often includes aliases for some entities. The alias lists are generally incomplete and aliases such as acronyms are highly ambiguous. So they can't be used as a definitive source for candidate generation. However, aliases can be used in `KRISS` to generate additional self-supervised mentions from unlabeled text, just like the preferred terms. To reduce noise, we similarly skip ambiguous aliases shared by multiple entities.

Semantic hierarchy Ontology often organizes entities in a hierarchy via `ISA` relationship among entities. For instance, in UMLS, the ER gene is assigned a *Semantic Tree Number* (A1.2.3.5), which specifies the `ISA` path from root to its entity type (*Gene or Genome*). For each entity in UMLS, we concatenate its semantic tree number (stn), entity type, as well as aliases to generate an *entity-centric reference* in the following form:

$$\texttt{[CLS]}\ \text{stn}\ \texttt{[SEP]}\ \text{type}\ \texttt{[SEP]}\ \text{aliases}\ \texttt{[SEP]}$$

We introduce a separate encoder to compute the vector representation $\mathbf{r}_e$ from the last-layer hidden state of [CLS] for entity e. For learning, besides the contextual vectors $\{\mathbf{c}_1, ..., \mathbf{c}_{2N}\}$ for N entities, a minibatch includes N entity-centric references $\{\mathbf{r}_{e_1}, ..., \mathbf{r}_{e_N}\}$. Given a positive pair $(\mathbf{c}_i, \mathbf{r}_{e_j})$, we treat the other $N-1$ entity-centric references as negatives and compute the InfoNCE loss:

$$\ell_{\mathbf{c}_i,\mathbf{r}_{e_j}} = -\log \frac{\exp(\mathbf{c}_i^\top \mathbf{r}_{e_j}/\pi)}{\sum_{k=1}^{N} \exp(\mathbf{c}_i^\top \mathbf{r}_{e_k}/\pi)}, \quad (4)$$

where π is a temperature parameter. The final loss between mentions and entity-centric references is computed across all positive pairs in a minibatch:

$$\mathcal{L}' = \frac{1}{2N} \sum_{k=1}^{N} [\ell_{\mathbf{c}_{2k-1},\mathbf{r}_{e_k}} + \ell_{\mathbf{c}_{2k},\mathbf{r}_{e_k}}] \quad (5)$$

We jointly optimize two contrastive losses $\alpha\mathcal{L} + \beta\mathcal{L}'$, with weights α and β. For inference, we include entity-centric references in $\texttt{Link}(m_q)$ as:

$$\texttt{Link}(m_q) = \operatorname*{argmax}_{e \in \mathcal{E}} \max_{m_p \in \texttt{Proto}(e)} \mathbf{c}_q^\top (\mathbf{c}_p + \mathbf{r}_e) \quad (6)$$

Entity description For a small fraction of common entities, manually written descriptions may be available. In UMLS, less than 6% of entities have description, so they can't be used as the main source for contrastive learning and linking. Still, the information may be useful and can be incorporated in KRISS by appending it to the entity-centric reference (separated by [SEP]).

3.6 Cross-Attention Candidate Re-Ranking

Inspired by Logeswaran et al. (2019); Wu et al. (2020), we further improve the linking accuracy by learning to re-rank the top K candidates via a *cross-attention* encoder. The input concatenates the mention and candidate representations (with the second [CLS] removed). A linear layer is applied to the top [CLS] encoding to compute the re-ranking score. The training data is generated by pairing self-supervised mentions with top K candidates based on $\texttt{Link}(m_t)$. We learn the encoder using a cross-entropy loss that maximizes the re-ranking score for the correct entity.

3.7 Lazy Learning

KRISS does not require labeled information in training or inference. However, if labeled examples are available, KRISS can directly use them, with zero additional training, as in *lazy learning* (Wettschereck et al., 2004). In this case, gold mention examples from target training data are used as mention prototypes for linking, augmenting the self-supervised ones. KRISS can also use labeled examples to fine-tune the self-supervised model, we leave it to future work.

	Mentions	Entities	Domain Entities
NCBI	6,892	790	16,317
BC5CDR-d	5,818	1,076	16,317
BC5CDR-c	4,409	1,164	233,632
ShARe	17,809	1,866	82,763
N2C2	13,609	3,791	423,670
MM (full)	352,496	34,724	3,416,210
MM (st21pv)	203,282	25,419	2,325,023

Table 1: Summary of entity linking datasets used in our evaluation. MM refers to MedMentions; st21pv refers to the subset with 21 most common semantic types. Domain entities refer to candidates in the UMLS sub-domains (e.g., disease) considered in the dataset.

4 Experiments

4.1 Entity Linking Benchmark

We consider seven standard entity linking datasets, spanning biomedical literature and clinical notes. See Table 1 for a summary. In particular, MedMentions (MM) (Mohan and Li, 2019) is the largest and most comprehensive dataset for biomedical entity linking, covering diverse UMLS entities (including all entity types in other datasets). See subsection A.2 for details. Training and development sets are not used in any way during self-supervised learning. Only test sets are used to evaluate self-supervised entity linking. we assume that gold mention boundaries are given and focus on evaluating linking accuracy. Given a test mention, the system needs to return the correct entity unique identifier to be considered as correct, as required by entity linking.

4.2 Implementation Details

For unlabeled text, we use the same corpus as in Gu et al. (2021), comprising 14 million PubMed abstracts. For domain ontology, we use UMLS 2017AA Active, containing 3.47 million entities.

We use a self-supervised dev set to choose hyperparameters. For self-supervised mentions, a mention-centered window of 64 tokens is used as context. We sample three mentions per entity for training, and sixteen as prototypes at test time. The encoders for mentions and entity-centric references are initialized with PubMedBERT (Gu et al.,

	NCBI	BC5CDR-d	BC5CDR-c	ShARe	N2C2	MM (full)	MM (st21pv)	Mean
QuickUMLS	39.7	47.5	34.9	42.1	29.8	12.1	20.0	32.3
BLINK	49.0	48.7	52.0	32.8	25.1	13.9	19.4	34.4
SapBERT†	63.0	83.6	96.2	80.4	59.7	37.6	44.2	66.4
KRISSBERT (self-supervised)	**83.2**±0.5	**85.5**±0.2	**96.5**±0.1	**84.0**±0.1	**67.8**±0.1	**61.4**±0.1	**63.5**±0.1	**77.4**
MedLinker	50.5	62.0	80.5	56.8	37.6	32.9	57.6	54.0
ScispaCy	66.8	64.0	85.3	66.6	54.6	53.1	52.9	63.3
KRISSBERT (supervised only)	76.9±0.9	85.5±0.7	93.8±0.3	53.9±0.4	29.2±1.2	60.7±0.3	63.7±0.4	66.2
KRISSBERT (lazy supervised)	**89.9**±0.1	**90.7**±0.1	**96.9**±0.1	**90.4**±0.1	**80.2**±0.1	**70.7**±0.1	**70.6**±0.1	**84.2**

Table 2: Comparison of test accuracy on standard entity linking datasets. Top four systems only use UMLS and unlabeled text. MedLinker and ScispaCy use MedMentions labeled examples for supervision. KRISSBERT (self-supervised) uses self-supervised mentions for learning and linking, whereas KRISSBERT (supervised only) uses training-set mentions instead. KRISSBERT (lazy supervised) augments KRISSBERT (self-supervised) with training-set mentions for linking, as in lazy learning (§3.7). **†SapBERT results are different from reported in Liu et al. (2021). We explain the difference in §4.5.**

2021). For learning, we use Adam with batch size 512, learning rate 10^{-5}, dropout rate 0.1, and both p_{mask} and p_{replace} 0.2. For simplicity, we set temperatures τ, π to 1.0, and loss weights α, β to 0.5. Training takes 3 hours on 4 NVIDIA V100 GPUs. We update parameters in all layers and denote the end model as KRISSBERT. At test time, we use FAISS (Johnson et al., 2019) with IndexFlatIP to obtain the top 100 prototypes for re-ranking. The inference speed is 5,461 mention queries per minute, on a test machine with Intel Xeon CPU E5-2690 and a Tesla P100 GPU.

4.3 Baseline Systems

We conduct head-to-head comparison against five baseline systems, including popular tools and prior state-of-the-art methods: QuickUMLS (Soldaini and Goharian, 2016), BLINK (Wu et al., 2020), SapBERT (Liu et al., 2021), MedLinker (Loureiro and Jorge, 2020), ScispaCy (Neumann et al., 2019). See subsection A.4 for details.

4.4 Main Results

Table 2 shows the main results. KRISSBERT results are averaged over three runs with different random seeds. As expected, QuickUMLS provides a reasonable dictionary-based baseline but can't effectively handle variations and ambiguities. BLINK attained promising results in the Wikipedia domain, but performed poorly in biomedical entity linking, due to the scarcity of available entity descriptions. SapBERT performed well on largely unambiguous entity types such as chemicals/drugs but faltered in more challenging datasets such as MedMentions. By contrast, KRISSBERT performed substantially better across the board, establishing new state of the art in self-supervised biomedical entity linking, outperforming prior best systems by 10 points in average and by over 20 points in MedMentions. The SapBERT results are different from Liu et al. (2021); we explain the difference in §4.5.

By leveraging knowledge-rich self-supervision, KRISSBERT even substantially outperformed supervised entity linkers such as MedLinker and ScispaCy, which used MedMention training data, gaining over 10-20 absolute points in average.

Self-supervised KRISSBERT also outperforms KRISSBERT (supervised only). It is particularly remarkable as KRISSBERT (self-supervised) learns a *single, unified* model for over three million UMLS entities, whereas KRISSBERT (supervised only) learns *separate* supervised models that tailor to individual datasets. This seemingly counter-intuitive result can be explained by *the unreasonable effectiveness of data* (Halevy et al., 2009). Knowledge-rich self-supervision produces a large dataset comprising diverse entity and mention examples. Despite the inherent noise, it confers significant advantage over supervised learning with small training data. This manifests most prominently in small clinical datasets like ShARe and N2C2.

4.5 Why the Entity Linking Scores Reported in the SapBERT Paper Are Incorrect?

The SapBERT paper (Liu et al., 2021) reported substantially higher scores than that in Table 2. Unfortunately, this stems from a significant error in their evaluation method, as inherited from BIOSYN (Sung et al., 2020) and widely adopted in subsequent work (e.g., Lai et al., 2021). Here, we conduct a detailed analysis using SapBERT (Liu et al., 2021) as the representative example.

The problem can be immediately discerned by first principle. SapBERT completely ignores the

Mention: "*... Hence, we aimed to find drug targets using the 2DE / **MS** proteomics study of a dexamethasone ...*"
SapBERT prediction: surface form MS, which is shared by multiple entities, such as Master of Science (C1513009), Mass Spectrometry (C0037813), etc.
`KRISSBERT` **prediction**: Mass Spectrometry (C0037813)
`KRISSBERT` **predicted prototype**: "*... **mass spectrometry** is a widely used technique for enrichment and sequencing of phosphopeptides ...*"

Example: "*... every patient followed up accordingly within ten days of **discharge** ...*"
SapBERT prediction: surface form DISCHARGE, which is shared by multiple entities, such as Discharge, Body Substance, Sample (C0600083), Patient Discharge (C0030685), etc.
`KRISSBERT` **prediction**: Patient Discharge (C0030685)
`KRISSBERT` **predicted prototype**: "*Performance of the Hendrich Fall Risk Model II in Patients **Discharged** from Rehabilitation Wards ...*"

Example: "*... 5 days of oral prednisone in treatment of adults with **mild** to moderate asthma exacerbations ...*"
SapBERT prediction: surface form MILD, which is shared by multiple entities, such as Mild Severity of Illness Code (C1547225), Mild Adverse Event (C1513302).
`KRISSBERT` **prediction**: Mild asthma (C0581124)
`KRISSBERT` **predicted prototype**:
"***Mild asthma** exacerbations in a group of children with cough as a dominant symptom ...*"

Example: "*... in patients with thyroid nodules evaluated as Bethesda Category III (**AUS** / FLUS) in cytology ...*"
SapBERT prediction: surface form AUS, which is used by Australia (C0004340).
`KRISSBERT` **prediction**: Atypical cells of undetermined significance (C0522580)
`KRISSBERT` **predicted prototype**:
"***Atypia of undetermined significance** (AUS) or follicular lesion of undetermined significance (FLUS), as stated by The Bethesda System for Reporting Thyroid Cytopathology ...*"

Table 3: Examples of ambiguous mentions: SapBERT struggles whereas `KRISSBERT` predicts correctly.

	Mention As-is	SapBERT	`KRISSBERT`
NCBI	76.9	92.0	91.3
BC5CDR-d	83.4	93.8	92.8
BC5CDR-c	92.3	96.5	97.2
ShARe	74.5	85.6	87.3
N2C2	61.2	67.9	76.1
MM (full)	47.1	52.2	71.3
MM (st21pv)	48.3	53.8	72.2
Mean	69.1	77.4	84.0

Table 4: Accuracy comparison based on the evaluation metric used by Liu et al. (2021).

	Ambiguous(%)	SapBERT	`KRISSBERT`
NCBI	43.2	57.1	64.5
BC5CDR-d	30.7	63.9	64.5
BC5CDR-c	11.5	76.4	76.5
ShARe	48.5	67.5	72.4
N2C2	67.5	50.7	58.2
MM (full)	67.8	24.8	48.9
MM (st21pv)	69.4	29.6	52.5

Table 5: Accuracy comparison on ambiguous cases.

	`KRISSBERT` (lazy supervised)	Supervised State of the Art
NCBI	89.9	89.1 (Ji et al., 2020)
BC5CDR	93.7	91.3 (Angell et al., 2021)
ShARe	90.4	91.1 (Ji et al., 2020)
N2C2	80.2	81.6 (Xu et al., 2020)
MM (full)	70.7	45.3† (Mohan and Li, 2019)
MM (st21pv)	70.6	74.1 (Angell et al., 2021)

Table 6: Comparison of test accuracy of `KRISSBERT` with lazy learning (§3.7) and supervised state of the art. †Prior work generally avoids evaluating on the full MM dataset; we can only find one published result which combines boundary detection and linking.

context of an entity mention (e.g., see Footnote 3 and Formal Definition in Section 2 in Liu et al., 2021). Given an ambiguous mention, there is no way such methods can resolve the ambiguity. Instead, these methods would merely produce a surface form (Footnote 2 in Liu et al., 2021). If the surface form matches multiple entities in name or alias, these methods can't predict a unique entity as required by entity linking. Unfortunately, such an ambiguous prediction is considered correct by their evaluation, as long as the gold entity is one of the matching entities.[3]

Table 3 shows examples of such ambiguous cases. E.g., given the mention "MS", without the context SapBERT has no way to resolve its ambiguity. Instead, it simply returns a verbatim surface form "MS", which can be mapped to many UMLS entities. Following BIOSYN, SapBERT evaluation would simply considers this as correct, as one of the matching entities is the gold entity Mass Spectrometry (C0037813). However, this obviously does not reflect the true linking performance for SapBERT, as it can't distinguish it from other equally matching entities such as such as Master of Science (C1513009) and Montserrat Island (C0026514).

Even if we adopt this incorrect evaluation method, `KRISSBERT` still substantially outperforms SapBERT, especially on the largest and most challenging MedMention dataset (see Table 4). The gain stems from cases when the gold entities have no official aliases matching the surface form predicted by SapBERT, whereas `KRISSBERT` can still match the gold entity based on context (e.g., see the last two examples in Table 3). We also evaluated the trivial baseline that returned the mention as is and found that SapBERT often does not out-

[3] See the code at `https://github.com/cambridgeltl/sapbert/blob/main/evaluation/utils.py#L42`

	NCBI	BC5CDR-d	BC5CDR-c	ShARe	N2C2	MM (full)	MM (st21pv)	**Mean**
KRISSBERT	**83.2**	**85.5**	**96.5**	**84.0**	**67.8**	**61.4**	**63.5**	**77.4**
– cross-attention re-ranking	82.8	85.0	95.1	83.4	65.0	59.4	61.3	76.0
– mention pair contrast	77.9	82.2	93.3	75.0	56.3	47.8	49.9	68.9
– aliases	83.2	85.2	96.4	84.0	67.7	61.0	63.2	77.2
– semantic hierarchy	82.7	85.1	96.4	83.0	65.7	59.0	61.5	76.3
– entity description	83.1	85.4	96.3	84.0	67.8	61.2	63.4	77.3
Initialize w. BERT	79.3	80.6	94.4	74.5	58.4	53.9	55.3	70.9

Table 7: Ablation study of KRISSBERT on the impact of knowledge components and domain-specific pretraining.

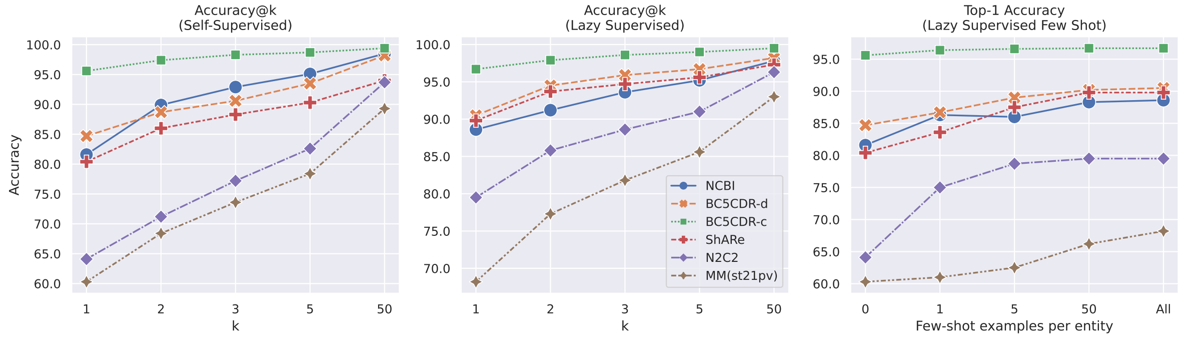


Figure 2: Test accuracy (oracle) with top K predictions shows that improving ranking has the potential to yield large gains. Few-shot learning results are averaged over three runs.

perform it by much, especially on the most representative MedMention dataset. Interestingly, under this inflated evaluation, KRISSBERT appears to slightly underperform SapBERT in the relatively easy datasets NCBI and BC5CDR-d (both about diseases). We found that, in rare occasions, the context may lead KRISSBERT to predict a more fine-grained concept (see subsection A.5).

As shown in Table 5, ambiguous mentions[4] abound, especially in more diverse and realistic datasets such as N2C2 and MedMentions. The SapBERT paper's evaluation thus reflects the oracle score (assuming that the right entity is always chosen out of multiple candidates), rather than true linking performance. For more realistic assessment, if SapBERT returns multiple entities, a random one would be chosen for evaluation, as in §4.4. Not surprisingly, KRISSBERT substantially outperforms SapBERT in the ambiguous cases, but still has much room for growth.

4.6 Lazy Supervised Entity Linking

KRISSBERT can make good use of labeled data when available. Even lazy learning (§3.7) yields results comparable to supervised state of the art, as shown in Table 6. Note that KRISSBERT (lazy supervised) is based on a single task-agnostic model (KRISSBERT (self-supervised)), and simply uses corresponding training set examples as prototypes for linking in a zero-shot fashion. By contrast, prior supervised state-of-the-art results were attained using separate models that tailored to individual datasets. They may use additional supervision such as coreference and joint inference (Angell et al., 2021), which can be incorporated into KRISSBERT.

4.7 Ablation Studies

In Table 7, we conduct a series of ablation studies to understand the impact of domain knowledge and model choices. Deep cross attention between query mentions and candidates produces consistent gains. The mention pair contrastive loss $\mathcal{L}$ (§3.2) is fundamental for self-supervised learning, whereas additional domain knowledge such as entity descriptions and semantic hierarchy offer incremental gains. Domain-specific pretraining (PubMedBERT; Gu et al., 2021) offers a substantial advantage for biomedical entity linking, gaining 6.5 points on average over BERT initialization.

4.8 Discussion

Aside from BC5CDR-c where KRISSBERT already performs very well, there is a large gap (10-15 points) between top-1 and top-5 accuracy, in both self-supervised and lazy supervised settings (Figure 2). This suggests that there is much room for KRISSBERT to gain by further improving ranking.

[4] We consider a mention as ambiguous if it can't be matched to a unique entity as is.

KRISSBERT also facilitates efficient few-shot learning, with a single example per entity yielding over 10 point gain in N2C2. subsection A.5 Table 9 shows examples of common errors by KRISSBERT. They are subtle and challenging. E.g., the gold concept is expression, while KRISSBERT predicts the procedure of expression.

5 Conclusion

We propose knowledge-rich self-supervised entity linking by conducting contrastive learning on mention examples generated from unlabeled text using available domain knowledge. Experiments on seven standard biomedical entity linking datasets show that our proposed KRISSBERT outperforms prior state of the art by as much as 20 points in accuracy. Future directions include: further improving self-supervision quality; incorporating additional knowledge; applications to other domains.

Limitations

KRISS is mainly tested for languages with limited morphology, i.e., English. Relatively large GPU resources, 4 NVIDIA V100 GPUs, are required to train the KRISSBERT model. Therefore, we did not do an exhaustive search for hyperparameters. Our experiments report results on seven standard biomedical datasets, which may not reflect KRISSBERT performance in the real-world applications.

References

Simon Almgren, Sean Pavlov, and Olof Mogren. 2016. Named entity recognition in Swedish health records with character-based deep bidirectional LSTMs. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 30–39, Osaka, Japan. The COLING 2016 Organizing Committee.

Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. Clustering-based inference for biomedical entity linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608, Online. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Seattle, Washington, USA. Association for Computational Linguistics.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

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

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.

Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. 2014. Modeling interestingness with deep neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2–13.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng

Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(02):8–12.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–34, Sofia, Bulgaria. Association for Computational Linguistics.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2333–2338, New York, NY, USA. Association for Computing Machinery.

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

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1631–1639, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

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. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016. Baw068.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. 2005. Rxnorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Daniel Loureiro and Alípio Mário Jorge. 2020. Medlinker: Medical entity linking with neural representations and dictionary matching. In *Advances in Information Retrieval*, pages 230–237, Cham. Springer International Publishing.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/UMass Lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Sunil Mohan and Donghui Li. 2019. Medmentions: a large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 97–109, Melbourne, Australia. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

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

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217, Boulder, Colorado. Association for Computational Linguistics.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland. Association for Computational Linguistics.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Dan Roth, Heng Ji, Ming-Wei Chang, and Taylor Cassidy. 2014. Wikification and beyond: The challenges of entity and concept grounding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials*, page 7, Baltimore, Maryland, USA. Association for Computational Linguistics.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 101–110.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, Michigan. Association for Computational Linguistics.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Dietrich Wettschereck, David W. Aha, and Takao Mohri. 2004. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11:273–314.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Dongfang Xu, Manoj Gopale, Jiacheng Zhang, Kris Brown, Edmon Begoli, and Steven Bethard. 2020. Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (bert)–based ranking for concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1510–1519.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In

3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Shaodian Zhang and Noémie Elhadad. 2013. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6):1088–1098. Special Section: Social Media Environments.

A Appendix

A.1 Additional Related Work

Zero-shot entity linking Recent work (Logeswaran et al., 2019) enables generalization to unseen entities by learning a cross-attention BERT model over the mention and entity contexts for candidate ranking. Gillick et al. (2019); Wu et al. (2020) introduce a bi-encoder that encodes the mention context and entity context separately, thus scaling to candidate generation and reducing recall loss due to mention variations. These methods, however, still require labeled information such as gold mention examples, which are not readily available in many high-value domains. This restricts their applicability to the Wikipedia domain, where labeled mentions can be gleaned from hyperlinks and entity pages. KRISSBERT, however, does not require labeled information and can learn from entity list and unlabeled text alone.

Contrastive learning Contrastive learning conducts representation learning by mapping semantically similar instances to nearby points (Hadsell et al., 2006). Contrastive loss is often a variant of noise-contrastive estimation (NCE) that normalizes against negative (dissimilar) examples (Gutmann and Hyvärinen, 2010). A popular choice is InfoNCE (Oord et al., 2018), where each minibatch samples a query instance (q), a few instances k_i's with one positive (similar) example $k+$, and optimizes the softmax of the query's dot product with the positive example $L(q) = -\log(\exp(q \cdot k+)/\sum_i \exp(q \cdot k_i))$. In computer vision, contrastive learning is often synonymous with self-supervised learning, where "similar" images are generated using data augmentation techniques assumed to preserve semantics (e.g., crop, resize, recolor) (Wu et al., 2018; Oord et al., 2018; He et al., 2019; Chen et al., 2020). In NLP, contrastive estimation has been applied to probabilistic unsupervised learning (by approximating the partition function with a tractable neighborhood) (Smith and Eisner, 2005; Poon et al., 2009). With the rise of neural representation, contrastive learning has also been applied to information retrieval (Huang et al., 2013; Shen et al., 2014), knowledge graph embedding (Bordes et al., 2013; Yang et al., 2015), entity linking (Loureiro and Jorge, 2020; Logeswaran et al., 2019; Wu et al., 2020), question answering (Karpukhin et al., 2020), typically with supervised labeled examples. In this paper, we apply contrastive learning to self-supervised entity linking where "similar" mentions are derived from unlabeled text using entity names and other domain knowledge, without requiring any labeled data.

A.2 Entity Linking Datasets

NCBI (Doğan et al., 2014) contains 793 PubMed abstracts annotated with 6892 disease mentions, which are mapped to 790 unique concepts in MeSH[5] or OMIM[6], both part of UMLS. BC5CDR (Li et al., 2016) contains 1,500 PubMed abstracts with 5,818 annotated disease mentions (BC5CDR-d) and 4,409 chemical mentions (BC5CDR-c), which are mapped to MeSH.

ShARe (Pradhan et al., 2014) contains 431 de-identified clinical reports with 17,809 disease mentions mapped to the SNOMED-CT (Spackman et al., 1997) subset of UMLS.

N2C2 (2019 n2c2/UMass Lowell shared task 3) (Luo et al., 2020) adds entity linking annotations to a subset of the 2010 i2b2/VA shared task dataset (Uzuner et al., 2011). The resulting dataset contains 100 de-identified discharge summaries with 13,609 mentions (including medical problems, treatments, and tests) linked to RxNorm (Liu et al., 2005) and SNOMED-CT (Spackman et al., 1997) within UMLS.

MedMentions (Mohan and Li, 2019) (MM) is the largest publicly available dataset for biomedical entity linking, which contains 4,392 PubMed abstracts and 350,000 mentions annotated with UMLS linking. MM (st21pv) is a sub-corpus limited to 21 most common entity types.

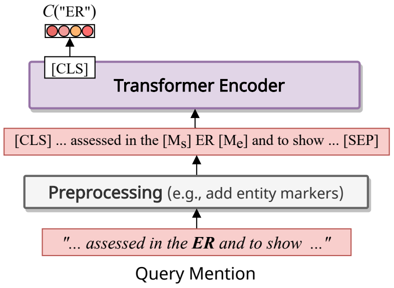
A.3 Contextual Mention Encoders

A.4 Baseline Systems

QuickUMLS (Soldaini and Goharian, 2016) conducts entity linking by approximate matching of mentions against UMLS entity lexicon (canonical name and aliases). It serves as a representative baseline for ontology-based entity linking.

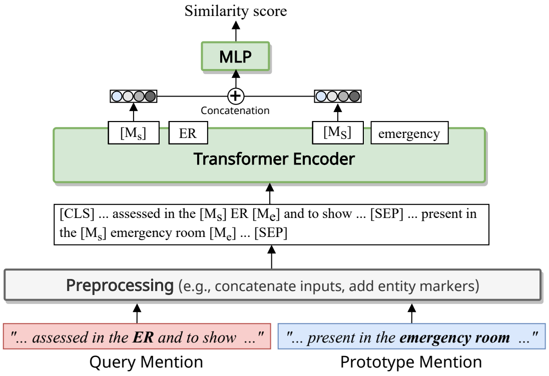
[5] www.nlm.nih.gov/mesh/meshhome.html

[6] omim.org



C("ER")
[CLS]
Transformer Encoder
[CLS] ... assessed in the [M$_s$] ER [M$_e$] and to show ... [SEP]
Preprocessing (e.g., add entity markers)
"... assessed in the ER and to show ..."
Query Mention

Figure 3: Contextual mention encoder for self-supervised entity linking.



Similarity score
MLP
Concatenation
[M$_s$]
ER
[M$_s$]
emergency
Transformer Encoder
[CLS] ... assessed in the [M$_s$] ER [M$_e$] and to show ... [SEP] ... present in the [M$_s$] emergency room [M$_e$] ... [SEP]
Preprocessing (e.g., concatenate inputs, add entity markers)
"... assessed in the ER and to show ..."
"... present in the emergency room ..."
Query Mention
Prototype Mention

Figure 4: Cross-attention candidate re-ranking.

Zero-shot entity linking by reading entity descriptions (Logeswaran et al., 2019; Wu et al., 2020) learns to encode contextual mentions against entity descriptions and attains state-of-the-art zero-shot entity linking results in the Wikipedia domain. Prior work uses gold mention examples in supervised learning. We adapt it to self-supervised learning using the self-supervised mention examples and available entity descriptions in UMLS. Prior work initializes the encoder with general-domain BERT models. To ensure head-to-head comparison, we followed KRISSBERT to use PubMedBERT (Gu et al., 2021) instead, which yielded better results.

SapBERT (Liu et al., 2021) learns to resolve variations in entity surface forms using synonyms in UMLS, using PubMedBERT (Gu et al., 2021). It ignores the mention context and returns all entities with a matching surface form. To use SapBERT for linking, we randomly select an entity when SapBERT returns multiple ones.

MedLinker (Loureiro and Jorge, 2020) is a strong *supervised entity linking* baseline that trains a BERT model on MedMentions. during test, it augments BERT-based prediction with approximate dictionary match for entities unseen in training.

ScispaCy (Neumann et al., 2019) provides another strong entity linking baseline that leverages labeled data in MedMentions to tune an elaborate biomedical linking system that uses TF-IDF based approximate matching and sophisticated abbreviation expansion.

A.5 Error Analysis

In Table 8, KRISSBERT considers "*t cell prolymphocytic leukemia*" and "*families with*" in the context of two mentions, and predicts more specific entities than the gold ones.

Mention: "*By analysing tumor DNA from patients with sporadic t cell prolymphocytic leukemia, a rare clonal malignancy with similarities to a **mature t cell leukemia** seen in ataxia telangiectasia ...*"
Gold entity: T-Cell Leukemia (C0023492)
KRISSBERT **prediction**: T-Cell Prolymphocytic Leukemia (C2363142)

Example: "*The majority (81%) of the breast ovarian cancer families were due to BRCA1, with most others (14%) due to BRCA2. Conversely, the majority of families with **female breast cancer** were due to BRCA2 (76%).*"
Gold entity: Breast cancer (C0006142)
KRISSBERT **prediction**: Familial cancer of breast (C0346153)

Table 8: Examples where KRISSBERT "misguided" by the context.

Mention: "*... NTeff cells appeared to have lower **expression** of Foxp1 ...*"
Gold entity: Protein Expression (C1171362)
KRISSBERT **prediction**: Expression Procedure (C0185117)
KRISSBERT **predicted prototype**: "*... **expression** of a myeloid differentiation antigen, Mo1 ...*"

Mention: "*... On admission included BUN / **creatinine** of 33/2.1 . Sodium 141 . ...*"
Gold entity: Creatinine Measurement (C0201975)
KRISSBERT **prediction**: Creatinine (C0010294)
KRISSBERT **predicted prototype**: "*... Sorbent binding of urea and **creatinine** in a Roux-Y intestinal segment. ...*"

Table 9: Examples of common errors by KRISSBERT.

A.6 License of Scientific Artifacts

UMLS (Bodenreider, 2004) is licensed to individuals for research purposes.[7] NCBI (Doğan et al., 2014) is under the terms of the United States Copyright Act.[8] BC5CDR is freely available for the research community.[9] ShARe (Pradhan et al., 2014) is under The PhysioNet Credentialed Health Data License.[10] N2C2 (Luo et al., 2020) is under the Data Use and Confidentiality Agreement.[11].

[7] uts.nlm.nih.gov/uts/assets/LicenseAgreement.pdf
[8] huggingface.co/datasets/ncbi_disease
[9] biocreative-v/track-3-cdr
[10] shareclefehealth2014task2/view-license/1.0
[11] n2c2.dbmi.hms.harvard.edu/data-use-agreement