OntolinkX: a context-aware linking approach integrating SapBERT and a cross-encoder reranker with hard negative training scenario for enhancing biomedical entity linking task

MSc in [Computing Science]
University of Glasgow

15th November 2024

Author: Gan Wang
Master of Science in [Computing Science]
at the University of Glasgow

Abstract

Biomedical Entity Linking (BEL), a crucial task in natural language processing, involves mapping mentions of biomedical entities in free text to their corresponding concepts in standardized and structured biomedical ontologies such as the Unified Medical Language System (UMLS). The increasing volume of biomedical literature and the complexity of medical terminologies present significant challenges for BEL, including entity ambiguity, dynamic knowledge bases, evolving terminology, and the need to maintain accuracy across diverse biomedical domain texts. Existing BEL systems often struggle with disambiguation, especially in the face of minimal context or sparse ontology descriptions, leading to reduced generalization ability in retrieval performance.

To address these challenges, we propose OntolinkX model, a context-aware linking approach that integrates SapBERT and a cross-encoder reranker using hard negative sampling scenarios. It builds on SapBERT which is a state-of-the-art entity linking approach that mainly focuses on synonym disambiguation and semantic alignment via contrastive learning but does not take full contexts into account. We show that adding a cross-encoder improves on SapBERT's performance in entity linking tasks. We explored the impact of incorporating additional information into the representation of both mention text and ontology concepts, two essential components in entity linking tasks. We start by taking entity names to represent ontology entries, then progressively augment the representations with semantic types and definitions. On the mention side, we incorporate contextual information from surrounding tokens within a dynamic window size. Furthermore, we examine the combined effect of full contextualized mention representations and enriched ontology representations.

Our two-stage pipeline begins with SapBERT retrieving potential entity candidates for each mention text. In the second stage, a cross-encoder is trained with negative sampling learning approach, starting from randomly generated negative samples and progressing to challenging "hard negatives", which are closest incorrect candidates from the retriever. Experiments show that incorporating richer information from both mention context and ontology descriptions improves retrieval performance. These findings suggest that our OntolinkX linking approach, alongside enriched representations from hard negative sampling strategy, can substantially improve BEL in complex biomedical texts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The volume of biomedical literature data such as published journal articles and clinical trial reports, has grown rapidly over the recent years. This literature contains valuable information about new discoveries and new insights that are continuously added to the already overwhelming quantity of literature [19]. Text mining tools can help researchers stay with the current latest discoveries by rapidly analysing those literature and identifying hidden connections between them [39]. This significantly boosts the speed of scientific knowledge discovery. Those tools automatically extract valuable information from vast amounts of biomedical records, which can aid the development in personalized treatment [38]. Furthermore, biomedical text mining helps assess potential interactions among drugs and-targets and flag adverse drug reactions by aggregating reports from different sources [5]. This simplifies complex search papers, making it easier for healthcare workers to grasp important insights. As a result, there is increasingly more demand for accurate biomedical text mining tools to extract meaningful information from this vast body of literature [34]. Among the key tasks in biomedical text mining are biomedical named entity disambiguation, biomedical relation extraction and biomedical question answering.

Biomedical entity linking, also known as named biomedical entity linking (BEL) or named entity disambiguation (NED), is a pivotal task in natural language processing (NLP) that involves identifying and disambiguating entities mentioned in text by linking them to a knowledge base, such as Wikipedia, DBpedia, or Unified Medical Language System (UMLS) which serves as a comprehensive knowledge base that unifies various health-related terminologies, enabling effective communication, data sharing, and information retrieval in the biomedical domain [4]. It can facilitate applications such as literature search, clinical decision making and relational knowledge discovery [23]. There are some substantial benefits of high-quality biomedical entity linking systems. Entity linking ensures clarity and accurate interpretation of information so that mentions of the same entity are consistently recognized across different sources [27]. Secondly, it enables the integration of data from multiple sources, which is crucial for clinical setting where disparate sources need to be merged to form a coherent knowledge base [32]. Furthermore, it improves search accuracy by bridging terms with their correct entities, improving information retrieval from large-scale databases [31]. These tangible benefits will bring about less confusion, less conflicting knowledge bases, less irrelevant retrieved results and even inaccurate and misleading outputs.

The entity linking problem can be addressed by framing it as a task of mapping

entity mentions to unified concepts in a medical knowledge graph as shown in Figure 1.1. The input to this diagram is a document with biomedical mentions within it. The right part of the figure is a knowledge base which contains a list of entities where each entity's detailed ontology information including name, sematic type and definition and so on are downloaded from UMLS. The major goal of BEL is to disambiguate each mention from a given document by linking it to an existing knowledge base [10].

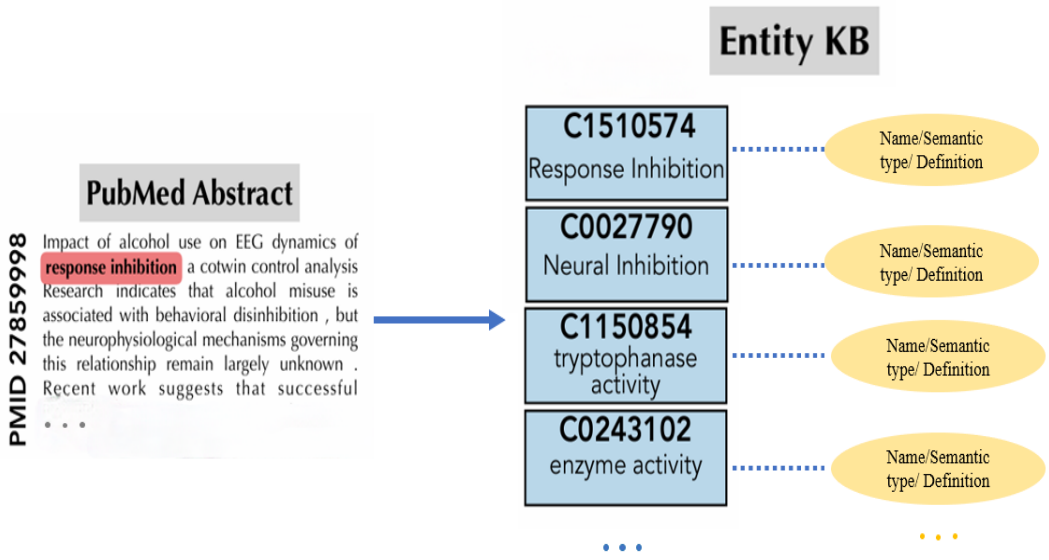

Figure 1.1: A common diagram of Biomedical entity linking task.

Current BEL faces several key challenges, including the dynamic nature of knowledge bases, the emergence of new entities, and the need for improved generalization ability, scalability and efficiency. Addressing these challenges requires ongoing research and innovation to further enhance the robustness and applicability of BEL systems in various NLP applications [22]. More specifically, BEL faces inherent difficulties such as ambiguity, where a single term can denote multiple entities based on context (e.g., "ADA" for "Adenosine Deaminase" or "American Diabetes Association"), and synonymy, where different terms can refer to the same entity (e.g., "myocardial infarction" and "heart attack"). Additionally, BEL methods have to contend with the dynamic terminology in biomedical domains, which continuously introduces new terms and concepts. Furthermore, achieving accurate entity linking requires a nuanced understanding of context, often necessitating information from the entire document rather than isolated sentences.

These challenges underscore the ongoing need for sophisticated approaches and technologies in BEL to effectively manage and interpret complex biomedical texts. Moreover, the creation of high-quality biomedical datasets is labour-intensive and resource-demanding. Many existing annotated datasets are limited in size, specificity, and less cross-domain [37], hindering the development of robust models that generalize well across diverse biomedical texts [40]. To a step further, the complexity and volume of biomedical literature require substantial computational resources for efficient processing and inference, posing scalability issues for entity linking algorithms [33].

Recent advancements in deep learning models [18], such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and more recently transformers,

have revolutionized the field of natural language processing (NLP). These models can capture complex patterns in data through multiple layers of nonlinear processing units, enabling them to learn hierarchical representations of text [29]. Many of them have been successfully applied to various biomedical NLP tasks, including named entity recognition (NER), relational extraction, and document classification, demonstrating significant performance improvements. Transformers, particularly models like BERT [8] and its biomedical variants (e.g., BioBERT [19]) have introduced a new paradigm for contextual understanding and scalability in entity linking. This contextual awareness is critical in biomedical texts where the meaning of entities can be highly dependent on surrounding information. These models leverage self-attention mechanisms to consider the full context of a sentence, which is essential for disambiguating terms that may have multiple meanings depending on their usage [19]. On the other hand, information retrieval (IR) has also contributed to many parts of different linking tasks. In the realm of IR, there are typically two main stages: retrieval, which generates an initial set of candidate entities, and reranking, where a more complex model is deployed to refine these results by evaluating the relevance of each candidate with respect to the query. It is notable that some leading methods for BEL do not follow this scenario. Therefore, we aim to incorporate the advantageous aspects of recent models and IR scenarios to enhance the BEL task.

To address above challenges, we propose a context-aware hybrid model named OntoLinkX, which integrates SapBERT [23] and a cross-encoder architecture as a linker to improve retrieval performance through re-ranking. This allows the model to capture the fine-grained semantic interactions between the query and candidate documents, making it more suitable for re-ranking tasks [28]. Using techniques from the field of IR, where re-ranking is often used to refine initial retrieval results, the cross-encoder model evaluates the relevance of each candidate document in relation to the query, ensuring more accurate and context-aware linking of biomedical entities. To further boost the model's generalization ability and scalability, we adopted a negative sampling scenario to better differentiate the most similar and tricky pairs. Our approach builds upon the success of existing deep learning models in biomedical NLP while introducing a more sophisticated re-ranking mechanism.

Chapter 2

Literature Review

2.1 Definitions and goals of Biomedical entity linking

BEL involves identifying and linking entities such as genes, proteins, diseases and drugs mentioned in biomedical texts to standardized entries in databases or ontologies [11]. This linking system helps unify diverse terminologies used across various biomedical domains including research, literature, clinical practice and studies. To take a specific instance, in the text of "Mutations in BRCA1 are linked to an increased risk of breast cancer," entity linking would involve associating entities "BRCA1" and "breast cancer" with databases like Entrez Gene or Disease Ontology, and generating possible candidates from them then disambiguating based on context to link "BRCA1" to the BRCA1 gene and "breast cancer" to the disease [13].

The primary inputs to BEL models are typically mention text, optionally to include the surrounding context (the windowed text), which is dependent on the window size parameter [7]. This window size refers to the number of tokens added from both the left and right of the mention text to form its final textual representation. In particular, a window size of k tokens surrounding the mention text could include relevant words that help clarify its meaning. In this scenario, it is typical to also add a [E] and [/E] to the left and right side of the mention text so that the model would have different focus between mention and its context. This setup helps distinguish between entities with similar names by capturing more clues about the mention's meaning in some particular contexts.

During BEL, the mention text is mapped to entity ontologies, which are possible matches from biomedical knowledge bases like UMLS. Candidate generation of entity ontology involves selecting potential matches based on lexical similarity, semantic closeness, or other matching techniques. UMLS, as one of the large biomedical knowledge bases, provides an aggregation of numerous large ontologies with standardized biomedical concepts [25]. Each entity is comprised of a canonical name (the standard name), synonyms that it is known by, a semantic type, a definition, and an identifier that uniquely tags it within the database. Including these elements may collectively enrich the linking process, helping to refine matches by providing comprehensive descriptors.

2.2 Commonly used datasets in BEL

Several datasets have become standard in the field of BEL and each of them offers distinct advantages for evaluating and developing BEL systems.

One of the most widely used datasets is MedMentions, which includes over 350,000 entity mentions linked to the UMLS. It is derived from over 4000 PubMed abstracts and supports tasks like mention disambiguation. The rich entity annotations and UMLS linkage make MedMention a prime choice for BEL tasks, particularly when targeting large-scale biomedical corpora [26].

Another commonly used dataset is BC5CDR (Biocreative V Chemical-Disease Relation) dataset, which focuses on identifying chemicals, diseases and their relations. It is particularly useful for extracting biomedical entities and has always been a key component in both NER and BEL evaluations [21].

Other popular datasets also play a crucial role in this field, such as NCBI (National Center for Biotechnology Information) Disease Corpus, OntoNotes and, BioCreative. The NCBI disease dataset includes disease mentions from PubMed abstracts linked to MESH (Medical Subject Headings) entries, providing a reliable foundation for disambiguating disease entities in biomedical text. For OntoNotes dataset, while not specific to biomedical domains, includes a broad range of entity annotations across different genres [30]. This pattern makes it useful for cross-domain entity linking tasks and it has been often used in more general EL experiments, including biomedical applications. Finally, the BioCreative dataset offers challenges focused on specific tasks like protein and gene normalization [14]. These challenges have generated many corpora that are regularly employed for BEL system evaluations, particularly in extracting relationships between biological entities.

The above-mentioned datasets are pivotal in the field of BEL, enabling BEL models to be measured and evaluated with regard to their accuracy, scalability, and ability to generalize across varied biomedical texts.

2.3 Problem Definition

Biomedical entity linking (BEL) involves mapping entity mention text in biomedical abstract or passage to standardized entities within a knowledge base, such as UMLS. Formally, let D denote a text corpus, which consists of a set of documents $D = \{d_1, d_2, \ldots, d_n\}$. Within each document d_i, there are a set of extracted mention texts denoted by $M = \{m_1, m_2, \ldots, m_k\}$. Each $m_j \in M$ is a biomedical terminology or phrase (such as drug, disease, protein or so on) and has a start and end coordinates for tagging its location relative to other mentions within the same document. Let C denote a knowledge base, where $C = \{c_1, c_2, \ldots, c_N\}$ represents a set of biomedical concepts (e.g., CUIs). The goal of BEL system is to identify or link the mention text appearing in each document d_i to a corresponding unique ontology concept c_i in the knowledge base, in our case, UMLS. Thus, we want to learn a function mapping $f : M \to C$ such that each mention m_i is linked to the correct concept c_i based on the context of m_i within d_i.

A common setup uses a two-stage pipeline including retrieval and re-ranking stage. Each of them serves unique goals as follows:

1. **Retrieval:** A set of candidate entities is retrieved from a knowledge base on the basis of textual mention m.

2. **Reranking:** The retrieved candidates are re-ranked based on our pre-trained cross-encoder model and the most appropriate match is selected.

2.4 Existing approaches

Historically, BEL systems were based on rule-based methods, dictionary lookups, and classical machine learning models, which all associate with feature engineering [20]. Rule-based systems use handcrafted rules and heuristics to identify and link entities. These systems rely on predefined patterns, regular expressions, and domain-specific rules to match text with entity names. They were among the first methods used for BEL and are still effective in certain scenarios, particularly when dealing with structured and predictable text. However, they struggle with the variability and complexity of natural language, especially in unstructured biomedical literature. Dictionary-based approaches use a predefined dictionary or lexicon of biomedical terms and their synonyms. The text is scanned for matches against this dictionary to identify and link entities. Dictionary-based approaches are a foundational technique in BEL, offering simplicity and efficiency [35]. However, their effectiveness is heavily dependent on the quality and comprehensiveness of the dictionary. They are often used in combination with other methods to improve performance. Classical machine learning approaches use supervised learning algorithms to train models on annotated datasets. Features are extracted from the text and used to train classifiers that predict entity links. Classical machine learning approaches brought significant improvements to BEL by leveraging statistical learning [17]. However, they require substantial feature engineering and annotated data, making them resource-intensive. Despite these challenges, they remain valuable, particularly when combined with modern techniques.

The advent of deep learning, particularly with the introduction of BERT (Bidirectional Encoder Representations from Transformers) [8] and its variants has revolutionized the field by providing powerful contextual embeddings that improve entity disambiguation. Models like BioBERT further optimized this framework by pre-training on large-scale biomedical corpora, significantly enhancing performance on BEL tasks. Thanks to the prior knowledge incorporation of pre-trained language models, recent state of the art BEL models have evolved quite rapidly, which are mainly in the leverage of deep learning and transformer-based pre-trained language models. For instance, one of the leading models in this domain is BiomedBERT (formerly known as PubmedBERT), which has shown large improvements in understanding biomedical texts due to its pre-training on large biomedical corpora [12]. Categorically, there are broadly three types of SOTA BEL models, namely Alias Matching EL, Contextualized EL and Auto regressive EL. The differences of aforementioned categories lie on their methodologies about how they handle semantic similarity, text matching, contextual reliance.

Alias Matching EL identify entities directly by comparing mention texts with known aliases in a knowledge base, thus providing a straightforward way for exact string match based on assessing the similarity between term embeddings. Some representative models in this type include SciSpacy [27], MetaMap [3], BioSyn [36] and SapBERT [23]. SciSpacy is complex biomedical text mining tool that evolved from the base NLP lib-

rary Spacy. It provides pipelines with pre-trained models for tasks like named entity recognition (NER) and entity linking. SciSpacy includes domain-specific vocabularies and supports linking entities to external biomedical databases like UMLS. It acts as a comprehensive toolkit, tailored for analyzing biomedical text using various linguistic and domain-specific resources. It also uses character n-grams for its entity linking, but integrates this with other SpaCy NLP capabilities, such as tokenization and syntactic parsing, to enhance the processing of biomedical data. MetaMap is one of the earliest tool for BEL, developed by the National Library of Medicine (NLM). It incorporated a combination of lexical and linguistic matching between input mention text and potential candidate terms from UMLS. The mapping search is computed based on string's similarity and its linguistic features. This model has made significant contributions on being effective for well-defined biomedical terms and is highly configurable to adapt to different biomedical subdomains. On the other hand, MetaMap lacks sufficient ability in understanding the surrounding contexts of mention text and heavily relies on the coverage and quality of UMLS. BioSyn is a synonym-based linking model for biomedical texts, which leverages each mention surface form to the best alias seen at training time to handle the nuanced terminology in biomedical literature. However, its effectiveness depends on computational resources and high-quality biomedical data. SapBERT is a transformer-based BEL model designed based on synonym-focused training objective to better handle complex biomedical terminologies with many aliases. It is built on BERT's architecture but trained specifically to identify and link biomedical terms using the UMLS database. During training, SapBERT pairs mention text with their corresponding synonyms from UMLS and is encouraged to produce similar embeddings for them while maximizing the distance between non-synonyms, improving the model's discriminative power on BEL tasks. Furthermore, it is built on BioMedBert that is pre-trained on large biomedical corpora like PubMed, allowing SapBERT to capture specific language of the biomedical domain. However, its effectiveness has no contextual understanding and is closely tied to the quality and comprehensiveness of its pre-training corpora.

Contextualized EL models, including MedLinker, ClusterEL, ArboEL [1], and KRISS-BERT, utilize embeddings derived from language models that take into account the context surrounding the entity mention for more accurate linking. MedLinker [24] and ClusterEL [2] use sophisticated neural network architectures to understand and disambiguate entities based on their contextual usage in the text. In contextualized BEL systems, there are two major essential architectures, namely Bi-encoders and Cross-encoders. They serve as efficient building blocks within aforementioned models for initial retrieval and final disambiguation. They play distinct yet complementary roles in contextualized models, enabling models to balance accuracy and efficiency according to the patterns of different datasets [11]. In a Bi-encoder setup, it generates embeddings for mention text and entities separately, typically with limited context understanding, and calculate similarities in a shared embedding space. For Cross-encoders, they are more context-sensitive as they encode mention-entity pairs together, providing deeper contextual integration while being computationally expensive. Their synergy has formed the foundation for multi-stage entity linking approaches. For ArboEL, it uses a hierarchical, multi-stage approach to entity linking. It starts with a bi-encoder to identify potential candidates and then refines them through hierarchical filtering based on entity types or ontology levels. Moreover, it incorporates a training

scheme to identify hard negatives, which led to better model precision in BEL tasks [1]. KRISSBERT uses a contrastive learning strategy on distantly supervised entity mentions. They show that this can be extended to a supervised setting without additional fine-tuning by simply swapping noisy prototypes for supervised ones, which achieves performance on-par with the best supervised EL models [42].

Autoregressive EL models like BioGenEL and BioBART treat entity linking as a sequence generation task, where the model generates the name or identifier of the entity token by token, conditioned on the input text and previously generated tokens [41]. While still in the experimental stages, autoregressive models show potential in handling rare and novel entities by dynamically generating entity representations based on context, thus overcoming some limitations of alias matching and contextualized models. However, these models are computationally intensive and require large amounts of training data to achieve competitive performance, which can be a significant barrier to their widespread adoption.

2.5 Research Questions

Given the challenges and progress of current approaches, Alias Matching EL models generally handles complicated jargon and various synonyms of the same concept well. For instance, SapBERT has demonstrated strong performance on several benchmarks, often outperforming other models [16], even without leveraging context. On the other hand, its major downside lies in no incorporation of sequence-level contextual understanding, which limits its capacity to disambiguate terms with similar or matching surface forms but different contexts (e.g. MS as 'multiple sclerosis' or 'mass spectrometry').

In light of the aforementioned aspects, this raises an important question: what if a context-aware reranker were integrated a leading alias matching method (i.e. SapBERT)? This would involve incorporating contextual understanding and re-ranking top retrieval entity candidates generated from SapBERT. To be more specific, we will explore and answer the four following **research questions.**

(1) RQ1: Will adding a cross-encoder based re-ranker after SapBERT retriever models improve entity linking in acc@k?

(2) RQ2: Could incorporating additional ontology information enhance its textual representation, thus better connect 'mention text' with its corresponding ontology?

(3) RQ3: Will adding a window of tokens around the "mention text" better link to its corresponding ontology?

(4) RQ4: Will hard negative sampling strategies help the model differentiate correct ontologies from wrong ones?

Chapter 3

Methodology

3.1 Dataset used for our experiments

The dataset we've used for conducting and comparing experiments is MedMentions. It currently has two released versions: full version and ST21pv subset versions. The full MedMentions is a very large-scale entity linking dataset containing over 4,000 abstracts and over 350,000 mentions across 200 semantic types linked to UMLS 2017AA. The ST21pv subset reduces this to a more manageable set of mentions (around 200,000), simplifying the data while preserving critical biomedical terminology. In our implementation, we adopted the ST21pv subset version.

The MedMentions ST21pv subset is a condensed version of full MedMentions, tailored specifically for high-priority biomedical entity types. This subset version selectively includes mention texts that fall into 21 semantic categories, which are chosen based on their frequency in biomedical literature. This version was created for the purpose of facilitating targeted research in biomedical NLP tasks like Named Entity Recognition (NER) and entity linking. By narrowing down the scope of semantic types to essential ones, the ST21pv subset simplifies data processing and allows models to concentrate on entities with the most biomedical relevance.

Most importantly, all entities in ST21pv subset are aligned to UMLS, which facilitates effective testing and evaluation of different BEL models as each mention text directly corresponds to an Standardized entry in the UMLS database. As a result, this version is well-suited for both training and evaluating BEL systems, making it more effective for conducting research in biomedical entity linking. The details of the dataset are shown below in Table 3.1.

Table 3.1: MedMentions Dataset Overview

Dataset Split	**Number of annotated Mentions**
Training	118,894
Validation	39,848
Testing	39,038
Total	**197,780**

3.2 Experimental setup

Data Preparation Details of UMLS. We firstly installed the full release of UMLS 2017AA version. We then extract all entity names from the MRCONSO.RRF raw file where duplicates are removed. Moreover, we extract both semantic types and definitions of entity ontology from MRSTY.RRF and MRDEF.RRF respectively, and add them as additional information for entities, which gives the model more understanding about the entity.

3.3 Baseline models

This section introduces the baseline models used to assess performance in BEL tasks. Each model provides a unique approach to embedding language representations, which allow us to leverage both general and domain-specific knowledge to improve entity linking and disambiguation in biomedical texts.

BERT (Bidirectional Encoder Representations from Transformers) [8] serves as foundational model for natural language processing. It is a pre-trained language model developed by Google that captures both left and right context in text. This is achieved through its training strategy of masked language modelling where random words in a sentence are hidden and predicted and next-sentence prediction where the model learns relations among sentences. These pretraining tasks equip BERT with a robust understanding of language structure and meaning, making it a robust baseline model for comparison in language understanding tasks, such as entity recognition and linking.

BioMedBERT [19] is a variant of BERT pre-trained specifically on PubMed abstracts and full-text articles, designed to improve performance on biomedical and clinical NLP tasks. It was originally known as PubMedBERT. This pretraining to biomedical texts enable BioMedBERT to capture terminologies, syntax and semantic relations within biomedical language, helping it outperform general-purpose BERT in biomedical NLP tasks, such as entity recognition. By learning domain-specific texts, it could generate embeddings that are more sensitive to nuances of biomedical terms, which enhances its potentials in BEL.

SapBERT [23], discussed previously, is a BERT-based model fine-tuned with self-alignment pretraining on UMLS (Unified Medical Language System) concepts, designed for biomedical entity linking. Its pretraining strategy learns to align synonymous entities from UMLS to a shared vector space, focusing on distinguishing the relations among similar or closely related biomedical terms. This is essential in BEL tasks where correct entity disambiguation are critical for accurate downstream applications in clinical settings.

3.4 Our model

We design a context aware two-stage learning framework that considers contextual information, synonym disambiguation and refining retrieved results for biomedical entity linking task. It consists of candidate retrieval using SapBERT embeddings and candidate reranking with a trained crossencoder using a negative sampling scenario. This pipeline enables efficient narrowing of potentially correct entities from large on-

tology knowledge base and further refines ranking based on their contextual relevance and resemblance.

3.4.1 Candidate Retrieval with SapBERT

In the initial stage, we represent both mention texts and entity ontologies in a shared vector space using SapBERT embeddings. Specifically, we denote a mention text and an entity ontology term as m and e respectively. Then we encode the mention m by SapBERT to get its vector representation $v_m = \text{SapBERT}(m)$. Each entity e in UMLS is also pre-encoded in the form of its vectorization as $v_e = \text{SapBERT}(e)$, where e contains only the entity name. To retrieve the candidate entities for m, we compute the cosine similarity between $\mathbf{v_m}$ and each vector $\mathbf{v_e}$ as:

$$\cos\_\text{sim}(\mathbf{v_m}, \mathbf{v_e}) = \frac{\mathbf{v_m} \cdot \mathbf{v_e}}{\|\mathbf{v_m}\|\|\mathbf{v_e}\|}$$

The top k entity candidates with the ranked highest similarity scores are retrieved, which serves to minimize the searching space during the reranking stage and maintain computationally effective. The detailed diagram of this whole procedure is shown below in Figure 3.1.

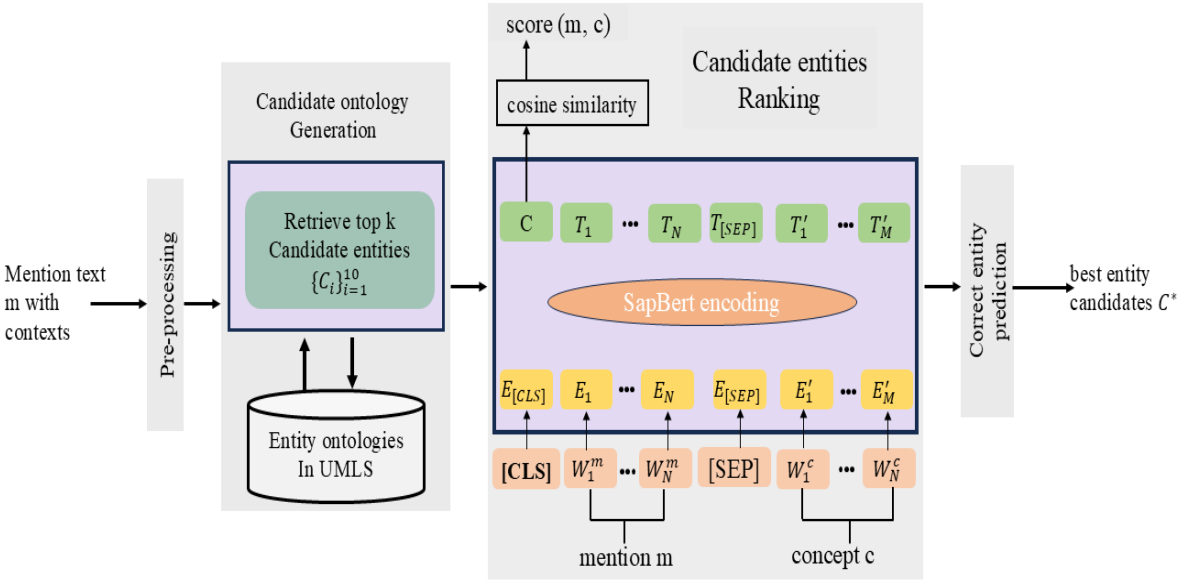

Figure 3.1: The retrieval stage using SapBERT in our whole pipeline.

3.4.2 Faiss: High-Dimensional Vector Search in BEL

Faiss [9] stands for Facebook AI Similarity Search and is an open-source library specifically designed for efficient similarity search. It is widely utilized in large-scale machine learning applications and deals with dense vector embedding in high-dimensional spaces. This library implements a variety of optimized algorithms that enable fast indexing and querying of vectors, such as product quantization, optimized GPU support etc [15]. These capabilities make Faiss particularly effective for tasks involving vast collections of embeddings within high retrieval speeds. In the context of BEL, we first initialize a Faiss index and store vectors of UMLS entities on it where each entity from UMLS is sent to SapBERT and we took the output [CLS] as its vector representation. This index is created for the nearest-neighbor search in the follow-up operation. When a query vector is sent to this Faiss index, it quickly retrieves the top k nearest entity vectors by computing the cosine similarity score, narrowing down the entity candidates for subsequent reranking process. This initial retrieval phase using Faiss reduces computational demands by largely filtering out irrelevant entities.

3.4.3 Candidate Reranking with Cross-Encoder

After we get the retrieved entity candidates in the first stage, we then refine them through a cross-encoder model f_θ, designed to assess and capture the semantic relevance between mention text along with its context and entity ontology pair. For a given mention text m, we denote a windowed text around the mention as its contextual information by context (m). The model encodes the mention text including its context as a vector

$$v_m = \text{SapBERT}(\text{left context}(m), [E]m[/E], \text{right context}(m))$$

where context (m) is the contextual background tokens surrounding the mention text from both left and right sides under a specific window size. $[E]$ and $[/E]$ are two special tags that we have used to highlight to the model that the content within middle of those two tags are the important input mention. We then treat the mention text m and entity candidate e (including its name, semantic type and definitions) as a single sentence

$$s = \text{concat}(\text{left context}(m), [E]m[/E], \text{right context}(m), [SEP], e)$$

where we put a [SEP] special token in the middle and take the whole sequence a sentence classification task. We model f_θ by BERT-based model and adopted the output [CLS] token regarded as the representation of the input. The cross-encoder, acting as binary classifier, outputs a probability score $P(y = 1 \mid s)$, representing the likelihood that e is the correct entity for m:

$$P(y = 1 \mid s) = \sigma(f_\theta(s))$$

where σ is the sigmoid activation function. Our cross-encoder is trained with binary cross-entropy loss:

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(P(y_i = 1 \mid s_i)) + (1 - y_i)\log(1 - P(y_i = 1 \mid s_i))]$$

where y_i is the ground truth label (1 if e is the correct entity, 0 otherwise) for the i-th mention-entity pair s_i. This loss function enables our model to learn discriminative features of both positive and negative data sample pairs.

3.4.4 Negative Sampling Scenario

The input to the cross-encoder model is an entity mention from a document paired with a candidate ontology term. To train the model, we created positive pairs from the MedMentions training data and used sampling techniques to generate negative pairs. In the initial setup, we generated randomly selected negative samples along with original positive samples to form a balanced training framework for cross-encoder reranker. Each random negative sample pair is comprised of mention text and a random entity ontology where this entity ontology must not be the corresponding correct entity of input mention. The number of random negative samples that we have generated is the same as the number of positive samples which equal to the quantity of mention texts in MedMentions datasets.We tested how difficult negative samples influence the model's decision making during inference. We pair up the mention text with incorrect candidate entities retrieved by SapBERT that are semantically similar or textually close to the ground-truth entity. This technique allows the cross-encoder reranker to be able to distinguish the minor differences between closely similar but incorrect entities with the authentic one. Formally, for a mention m, the set of hard negative samples $N_{\text{hard}}(m)$ is defined as:

$$N_{\text{hard}}(m) = \{e \in C \setminus c \mid \cos\_\text{sim}(\mathbf{v_m}, \mathbf{v_e}) \text{ is high}\}$$

where C represents all the entity ontologies in UMLS and c is the correct entity for the given mention m. The hard negative samples may help push the cross-encoder model to assign low similarity score for these incorrect but highly confusing entities during training, thereby improving the model's discriminative understanding ability.

3.4.5 Model training details

During training, we use SGD [6] with a learning rate of $3e{-}4$ to update the trainable parameters within the model. It is trained on the prepared MedMentions data samples with a batch size of 52 and with a range of 10-20 epochs. For early stopping, we monitor the model's performance on the validation set where we take the loss value of every epoch as the indicator of whether the model is trained towards the expected direction. If the validation loss stops decreasing for 3 epochs, the training is stopped early. This helps ensure the model generalizes well to new data by avoiding excessive training that could lead to overfitting on the training data. For the maximum length of each data sample pair, we set up the full length to 50 tokens as it is enough to cover the whole textual sequence of vast majority of data samples. Finally, the implementation is done under Python version 3.10 with PyTorch 2.2.1 and it takes approximately 6 hours on our machine for each experiment.

3.4.6 Performance Metric

BEL evaluation aims to assess the performance of EL systems in correctly identifying and linking entity mentions in text to their corresponding entries in a knowledge

base or ontology. The evaluation process typically involves measuring how accurately and completely the system can identify entities and link them to the correct entries in a given reference or gold standard dataset.

Accuracy@k measures the proportion of cases where the correct entity is among the top-k ranked candidates returned by the system. Its mathematical equation is defined as below:

$$\text{Acc@k} = \frac{N_k}{N}$$

where:

- N_k = Number of instances where the correct entity appears in the top k candidates.
- N = Total number of instances evaluated.

3.4.7 Inference and Evaluation

During inference stage, for a given mention m, we first encode its vector representation $\mathbf{v_m}$ using SapBERT. Then we compute cosine similarity of this query vector with all entities from UMLS to retrieve top k potential candidates. After that, each candidate e is paired up with the input mention text and passed through our pretrained cross-encoder reranker. The model will output a relevance score and recommend the linking to entities with the highest score:

$$\hat{c} = \arg \max_{e \in \{e_1, \ldots, e_k\}} P(y = 1 \mid \text{concat}(\text{ left context}(m), [E]m[/E], \text{ right context}(m), [SEP], e))$$

We evaluate the model's performance using top-k accuracy metrics, more specifically acc@1 and acc@5. These metrics reflect the frequency with which the correct entity is ranked first or among the top five candidates, respectively. It demonstrates the model's ability in linking biomedical mention text to its corresponding accurate UMLS concepts.

Chapter 4

Results

4.1 RQ1: Will adding a cross-encoder based re-ranker after SapBERT retriever models improve entity linking in acc@k?

For the first experimental setup, we explored how a cross-encoder based re-ranker would impact the entity linking system. Firstly, we employ SapBERT to generate dense embeddings for all biomedical entities in UMLS. Specifically, each entity is passed through the SapBERT model to obtain its vector representation. Here, we used the output [CLS] as the final representation of inputs. These embeddings are subsequently stored in a FAISS index, a library designed for efficient similarity search in high-dimensional spaces.

In our task, each mention text (from MedMentions dataset) is also encoded using SapBERT to obtain its corresponding embedding. We then calculate the cosine similarity between the query embedding and the pre-indexed entity embeddings to assess their similarity score by FAISS. The candidate entities with the highest similarity scores are retrieved, and the top-k candidates are ranked based on this score. Many approaches stop at this stage, without applying a reranker, and compute the acc@k to get the final performance. However, in our scenario, we expand one step further beyond this procedure which is to retrieve k+n more candidate entities instead of k. We then deploy the pre-trained cross-encoder model to re-rank k+n potential entity candidates to get top k ones. After that, we calculate the acc@k performance. Similarly, Bert, Pubmed-Bert and Sap-Bert have also been applied using the same strategy. The results are shown below in Table 4.1 and Table 4.2. The (5) and (10) on the Table 4.1 represent we retrieved 5 and 10 potential entity candidates in both tasks waiting for the re-ranking process, respectively.

Models	acc@1	acc@5
Bert	29.3	33.8
Pubmed-Bert	35.3	41.2
Sap-Bert	54.9	71.6
Reranker model A (only mention text – only entity name)	60.1 (5)	74.6 (10)

Table 4.1: Comparison between three baselines and our re-ranker model A

Retrieved entity candidates	1	5	10
Reranker model A (only mention text – only entity name)	54.9	60.1	67.0

Table 4.2: The effect of the number of retrieved candidates on final acc@1

Based on the results, we could observe that our re-ranker model A achieved relatively better performance on both tasks than SapBERT with roughly 5% higher in acc@1 and 3% higher in acc@5. Given the results, we believe the larger improvement for acc@1 is caused by two reasons. Firstly, finding the top 1 correct entity candidate is a slightly more difficult task than the top 5. In consequence, every model got lower values in acc@1 than acc@5. Second of all, cross-encoder based re-ranking process does help the BEL pipeline with some slight improvements as it increases the chance of finding the most correct entity by introducing more potential candidates, which is consistently reflected by the result of Table 4.2. However, this extra re-ranking process does add more computational and algorithmic complexity on the whole pipeline.

4.2 RQ2: Could incorporating additional ontology information enhance its textual representation, thus better connect 'mention text' with its corresponding ontology?

In the second experimental setup, we examined whether including additional ontology information from the ontology would have an impact on its bridging with mention texts. More specifically, we have adapted four different training information to be included for either mention text or entity ontology or for both. In detail, model A represents only "mention text" and only entity name from UMLS have been paired as samples to train the cross-encoder reranker. Model B takes "mention text" and additional ontology information including semantic type and definitions along with its name to train the cross-encoder. For the current setup, all results are shown below on Table 4.3. The other two training scenarios will be explained in detail later in RQ3 part.

Models	acc@1 (retrieve 5)	acc@1 (retrieve 10)	acc@5 (retrieve 10)	acc@5 (retrieve 15)
SapBERT	54.9	54.9	71.6	71.6
Reranker model A (only mention text – only entity name)	60.1	67.0	74.6	78.4
Reranker model B (only mention text – full entity information)	66.0	67.2	76.3	78.4

Table 4.3: Comparison between re-ranker model B with additional ontology information and re-ranker model A with only entity name and baseline SapBERT

We observe that re-ranker model B has achieved the best performance on both acc@1 and acc@5, which indicates adding more ontology information for entity candidates into the textual representation for mention text would boost the model's understanding about the relevance between two inputs (mention text and entity ontology). However, the improvement on metric acc@5 is relatively smaller compared with acc@1. Another clue is reranker model B achieves similar performance on acc@1 as model A when we increased the number of retrieved entity candidates from 5 to 10. It implies adding more information from the ontology is particularly useful when the number of retrieved entity candidates for re-ranking process is small. When the retrieved candidate set is larger, the model remains relatively unchanged.

4.3 RQ3: Will adding a window of tokens around the "mention text" better link to its corresponding ontology?

In the third experimental setup, we observe whether contextualized mention text would help the model better understand the relatively dynamic meaning of mention text, thus improve the entity linking system. Contextualized mention text involves including a window of tokens around the mention (e.g. "patients were given [E]aspirin[/E] to take for"). Apart from the trained re-ranker model A and model B in the previous experiment, we trained re-ranker model C and model D in this part. Regarding model C, we take the contextual information of the mention text to represent the textual mention while only keeping entity name as its ontology text. And we set up the window size of 15 to incorporate the contextual information for the mention text during training model C. For model D, we take mention text with contexts and additional ontology information together to train this cross-encoder re-ranker. For this setup, all results are shown below on Table 4.4

Models	acc@1 (retrieve 5)	acc@1 (retrieve 10)	acc@5 (retrieve 10)	acc@5 (retrieve 15)
SapBERT	54.9	54.9	71.6	71.6
Reranker model A (only mention text – only entity name)	60.1	67.0	74.6	78.4
Reranker model B (only mention text – full entity information)	66.0	67.2	76.3	78.4
Reranker model C (full mention text with context – only entity name)	66.2	68.7	76.4	78.5
Reranker model D (full mention text with context – full entity information)	63.5	64.7	76.3	78.3

Table 4.4: Comparison between four different reranker models and baseline SapBERT

Observing the results from Table 4.4, model C (based on contextualized mention text) outperformed any other models and improved 0.2% and 1.5% on acc@1 when respectively retrieving 5 and 10 potential entity candidates beyond the second best model B. This implies that contextualized information either for mention text or entity ontology would contribute more than simply adding more entity information from the ontology.

Another observation is the comparison between model A, model B and model C in Table 4.4. Both model B and C have got nearly the same performance on acc@5 while model A only catches up with them when the number of retrieved entity candidates is large enough to cover the correct ones. It indicates that it is possible for models that are trained with limited contextual and additional information to achieve relatively good performance as other more complex counterparts, and it is likely to be the case when the searching pool for potential entity candidates is big enough. As a result, people could choose to setup their model scenarios between incorporating more information to the model and increasing the retrieved potential candidates based on their specific experimental cases.

Lastly, by comparing the four columns of Table 4.4, it indicates that all models do get sequentially better scores on both metrics with an increased number of retrieved potential entity candidates. Model D outperformed model A on both acc@1 and acc@5 metrics given 10 retrieved entity candidates. However, model A beat back model D when we expand the re-ranking search by increasing the retrieved count to 15. Our suspicion is that adding much information including contexts either on mention text or entity ontology would make the model better understand the entity linking task. However, too much information may lead to noise that could overwhelm the model as

different models have their maximum extent or limitations of complexity in figuring out the entity linking task.

4.4 RQ4: Will hard negative sampling strategies help the model differentiate correct ontologies from wrong ones?

Lastly, we explore how different negative sampling strategies impact the ability of our trained cross-encoder re-rankers. We compared randomly negative sampling scenario with hard negative sampling approach. Since we have observed that re-ranker model C (which is trained based on contextualized mention text and entity name) generally outperformed any other models in previous experimental settings, in consequence, we adopted the same training information of model C for training the hard negative sampling model. Here, we trained the last two re-ranker models (model D and model E). For model E, we firstly generated the hard negative samples based on the SapBERT retriever. Then we combine them with the original positive samples together and shuffle all of them to train the cross-encoder to get the model E where the ratio of negative ones vs positive samples is 3 : 1. The results are shown below in Table 4.5.

We observe that model E achieved the best performance among all models in acc@1 with 0.4% and 0.3% higher than model C when retrieving 5 and 10 entity candidates. This is because the model E retrieved more correct results on some mention texts linking task on which previously model C sometimes retrieve wrong entity candidates (hard negative entities) that are closely similar to the ground-truth entity ontology. In addition, it has a small difference in acc@5 with model C. It suggests that hard negative samples make the cross-encoder re-ranker model slightly better understand the small patterns among most similar retrieved entity candidates, thus improving the performance after re-ranking process.

Models	acc@1 (retrieve 5)	acc@1 (retrieve 10)	acc@5 (retrieve 10)	acc@5 (retrieve 15)
Reranker model A (only mention text – only entity name)	60.1	67.0	74.6	78.4
Reranker model B (only mention text – full entity information)	66.0	67.2	76.3	78.4
Reranker model C (full mention text with context – only entity name)	66.2	68.7	76.4	78.5
Reranker model D (full mention text with context – full entity information)	63.5	64.7	76.3	78.3
Reranker model E (hard negatives) (full mention text with context – only entity name)	66.6	69.0	76.3	78.3

Table 4.5: Comparison of re-ranker models on different negative sampling strategies

Another observation is by comparing re-ranker model B, model D and model E on Table 4.5, it suggests that the performance of different re-ranker models regarding acc@5 is minor compared to their gaps in acc@1. This finding is consistent with the observation from RQ1 that finding the top 1 entity candidate appears to be a more difficult task than top 5 so that changes either in the model or training would have a larger impact in acc@1 than acc@5.

4.5 Discussion

The experimental results illustrate that our proposed model, which incorporates a cross-encoder reranker following SapBERT retriever and takes contextual information into account along with using hard negative samples, substantially enhances the accuracy of biomedical entity linking task. The specific performance on acc@1 and acc@5 show improvements over existing baseline models. These enhancements can be attributed to the combinatorial impacts of leveraging synonym disambiguation, semantic representation and contextual understanding afforded by our two-stage pipelines using hard negative sampling strategy.

In addressing our research questions, we find some evidence supporting the hypothesis that the incorporation of additional ontology information and contextual features could lead to more precise and accurate entity linking. The pair of mention text with entity ontology where either mention text takes its contexts or entity ontology comprising entity name, semantic type, and definitions provides a richer understanding for

the model to facilitate accurate linking.

Another critical component of our framework is the implementation of generating difficult, similar yet incorrect negative samples. By selecting challenging incorrect entities retrieved by SapBERT during training, we have equipped the cross-encoder reranker with differentiation capability among semantically and contextually similar samples. Results on the last table has demonstrated this is an effective strategy and suggested that the model has gained a robust ability to identify relevant entities while minimizing false positives.

Overall, the findings underscore the importance of advanced entity linking techniques in improving biomedical text mining and information retrieval. Biomedical entity linking have far-reaching implications, facilitating better knowledge extraction, and supporting clinical decision-making.

Chapter 5

Conclusion

In conclusion, we proposed a robust learning framework for biomedical entity linking task, which employs SapBERT for candidate retrieval and a pre-trained cross-encoder for raranking. This approach effectively enhances the model's ability to accurately disambiguate entities in complex biomedical texts. As demonstrated in our experimental evaluations, adding a cross-encoder reranker would substantially improve the model's performance than just synonym-focused SapBERT. Moreover, incorporating additional contextual information either for mention text or entity ontology can contribute some improvements for the whole BEL pipeline. Lastly, hard negative sampling strategy also helps model better understand the different patterns of data samples especially the ones that share similar contexts, thus enhancing the the model's ability for BEL. This two-stage pipeline's success is largely due to the integration of contextual understanding for mention text and the use of hard negative sampling in the training process. Together, these techniques equip the model with an optimized capability to distinguish closely relevant entities, a crucial factor in achieving higher accuracy in BEL.

As the volume of biomedical texts continues to expand rapidly, the accuracy and efficiency of BEL become increasingly critical. Our research contributes to the ongoing efforts to refine these techniques by providing a framework that not only enhances BEL accuracy but also being a comparable foundation for future advancements in biomedical text mining.

5.1 Limitations of our approach

Despite its promising results, our approach is not without its limitations. One primary concern is the reliance on the MedMentions dataset. It may be biased, particularly in the representation of biomedical entities, as it may focus on biomedical entities that are not relevant for other biomedical text mining tasks. While it is the largest dataset for entity linking, our experiments and their conclusions may not generalize across diverse biomedical texts. Expanding the diversity of datasets in future work may increase the model's applicability to a wider array of biomedical tasks.

In addition, though our method showed promising accuracy performance on the entity linking task, the computational resources required to apply this framework into large scale applications may face challenges for real-time implementation, particularly in other domains beyond biomedical ones.

5.2 Future Work

In the future, we would focus on expanding the proposed framework to address these limitations. Alternative candidate retrieval methods, such as coreference-based methods, could be explored to further optimize the initial entity candidate selection. Experimenting with different vectorization models and architectures for the cross-encoder could also enhance the whole pipeline's accuracy and adaptability.

Exploring innovative sampling techniques, particularly for negatives, may further improve the discriminative power of the model for closely similar entities. Moreover, incorporating broader and more specialized biomedical vocabularies could make our model become applicable across various biomedical subdomains, enhancing its exclusive values in fields, such as genomics, clinical diagnostics and pharmacology.

The significance of accurate biomedical entity linking cannot be overstated, as it plays a crucial role in enhancing numerous applications, including clinical decision making, facilitating biomedical knowledge discovery and so on. In consequence, continued innovation in BEL will be essential for maximizing the tangible benefits that biomedical data could bring to us and ultimately improving outcomes in healthcare and research.

Bibliography

[1] Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[2] Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. Clustering-based inference for biomedical entity linking. *arXiv preprint arXiv:2010.11253*, 2020.

[3] Alan R Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

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

[5] Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319, 2021.

[6] Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer, 2012.

[7] Hyejin Cho and Hyunju Lee. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20:1–11, 2019.

[8] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[10] Christina Du, Kashyap Popat, Louis Martin, and Fabio Petroni. Entity tagging: Extracting entities in text without mention supervision. *arXiv preprint arXiv:2209.06148*, 2022.

[11] Evan French and Bridget T McInnes. An overview of biomedical entity linking throughout the years. *Journal of biomedical informatics*, 137:104252, 2023.

[12] 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, 2020.

[13] Ming-Siang Huang, Po-Ting Lai, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*, 21(6):2219–2238, 2020.

[14] Rezarta Islamaj, Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Tiago Almeida, Richard AA Jonker, Sofia IR Conceição, Diana F Sousa, Cong-Phuoc Phan, Jung-Hsien Chiang, et al. The overview of the biored (biomedical relation extraction dataset) track at biocreative viii. *Database*, 2024:baae069, 2024.

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

[16] David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie S Mitchell. A comprehensive evaluation of biomedical entity linking models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 14462. NIH Public Access, 2023.

[17] John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA, 2001.

[18] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[19] 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.

[20] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. Cnn-based ranking for biomedical entity normalization. *BMC bioinformatics*, 18:79–86, 2017.

[21] 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.

[22] Yangning Li, Jiaoyan Chen, Yinghui Li, Tianyu Yu, Xi Chen, and Hai-Tao Zheng. Embracing ambiguity: Improving similarity-oriented tasks with contextual synonym knowledge. *Neurocomputing*, 555:126583, 2023.

[23] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*, 2020.

[24] Daniel Loureiro and Alípio Mário Jorge. Medlinker: Medical entity linking with neural representations and dictionary matching. In *European Conference on Information Retrieval*, pages 230–237. Springer, 2020.

[25] Nicolas Matentzoglu, Damien Goutte-Gattat, Shawn Zheng Kai Tan, James P Balhoff, Seth Carbon, Anita R Caron, William D Duncan, Joe E Flack, Melissa Haendel, Nomi L Harris, et al. Ontology development kit: a toolkit for building, maintaining and standardizing biomedical ontologies. *Database*, 2022:baac087, 2022.

[26] Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*, 2019.

[27] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

[28] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[29] N Patwardhan, S Marrone, and C Sansone. Transformers in the real world: A survey on nlp applications. information, 14 (4), 242, 2023.

[30] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, 2013.

[31] Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. Entity linking meets deep learning: Techniques and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2556–2578, 2021.

[32] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2014.

[33] Jiyun Shi, Zhimeng Yuan, Wenxuan Guo, Chen Ma, Jiehao Chen, and Meihui Zhang. Knowledge-graph-enabled biomedical entity linking: a survey. *World Wide Web*, 26(5):2593–2622, 2023.

[34] Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. Biomegatron: larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, 2020.

[35] Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics*, 22(6):bbab282, 2021.

[36] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*, 2020.

[37] Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. Cross-domain data integration for named entity disambiguation in biomedical text. *arXiv preprint arXiv:2110.08228*, 2021.

[38] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49, 2018.

[39] Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, and Zhiyong Lu. tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics*, 34(1):80–87, 2018.

[40] X Yang, A Chen, N PourNejatian, HC Shin, KE Smith, C Parisien, et al. A large language model for electronic health records. npj digit med. 2022; 5: 194.

[41] Hongyi Yuan, Zheng Yuan, and Sheng Yu. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. *arXiv preprint arXiv:2204.05164*, 2022.

[42] Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Knowledge-rich self-supervised entity linking. *arXiv preprint arXiv:2112.07887*, 2021.