World Wide Web (2023) 26:2593–2622
https://doi.org/10.1007/s11280-023-01144-4

Knowledge-graph-enabled biomedical entity linking: a survey

Jiyun Shi[1] · Zhimeng Yuan[1] · Wenxuan Guo[1] · Chen Ma[2] · Jiehao Chen[3] · Meihui Zhang[1]

Received: 26 October 2022 / Revised: 30 December 2022 / Accepted: 20 January 2023 / Published online: 2 May 2023

Abstract
Biomedical Entity Linking (BM-EL) task, which aims to match biomedical mentions in articles to entities in a certain knowledge base (e.g., the Unified Medical Language System), draws dramatic attention in recent years. BM-EL can help to disambiguate medical terms and link to rich semantic information in the biomedical knowledge base, which can act as an essential means for many downstream applications. Although entity linking tasks have been investigated in the general domain and achieved great success, many challenges remain in the biomedical field, for instance, highly complex terminology, less training data, and entity ambiguity. In this survey, we categorize BM-EL methods into rule-based, machine learning, and deep learning models according to the development of the model paradigm and provide a comprehensive review of each approach. In-depth study of current BM-EL efforts, we group the model architectures into four categories: joint entity recognition and linking, graph-based global entity disambiguation, cross-lingual architectures, and model-efficiency improvement. We further introduce six well-established datasets that are commonly used for BM-EL tasks. Furthermore, we present a comparison of the different methods and discuss their advantages and disadvantages. Finally, we discuss the limitations of existing methods for BM-EL and discuss promising future research directions.

Keywords Biomedical entity linking · Biomedical entity disambiguation · Knowledge base

1 Introduction

1.1 Motivation

Significant advancements in the healthcare field have led to the substantial growth of the biomedical corpus, including biomedical literature, electronic medical reports (EMRs), etc.

This article belongs to the Topical Collection: *Special Issue on Knowledge-Graph-Enabled Methods and Applications for the Future Web*
Guest Editors: Xin Wang, Jeff Pan, Qingpeng Zhang, Yuan-Fang Li

✉ Meihui Zhang
meihui_zhang@bit.edu.cn

Extended author information available on the last page of the article.

[1]. Mining and using the vital information in the biomedical corpus are highly helpful for researchers to build effective biomedical computational models for downstream biomedical analytics tasks. However, various problems with medical corpora, such as ambiguous meanings, abbreviations, misspellings, and missing text, have posed huge challenges in understanding medical texts. Therefore, it is important to disambiguate these terms by matching them to their corresponding concepts in the knowledge base, which promotes the study of Biomedical Entity Linking (BM-EL).

Knowledge Graph also known as Knowledge Base, represents a network of real-world entities, i.e. objects, events, situations, or concepts, and illustrates the relationship between them. BM-EL is proposed to match mentions in biomedical articles to entities in a biomedical knowledge base, e.g., the Unified Medical Language System (UMLS) [2]. Figure 1 shows an illustration of the BM-EL task. Its ability to link and exploit the rich semantic information in the biomedical knowledge base, entity linking (EL) acts as an essential means for many downstream applications, such as population and health analytics [3], medical information retrieval, question answering [4], and knowledge-graph construction. BM-EL task is one of the most important tasks in the domain of Biomedical Knowledge Graph research.

At present, most existing surveys of EL methods [5–9] focus on the general domain, and there are few discussions about the EL efforts in the biomedical domain. Considering that biomedical corpora contain more ambiguity and morphological variations than the corpora in the general domain, we argue that achieving the BM-EL task requires more analysis of the characteristics of the biomedical field. To this end, we pay attention to the rapidly developing BM-EL in this survey. We review the technology development for the BM-EL task and present the discussion about their characteristics and limitations.

In this survey, we review the papers published for the BM-EL task in the past, summarize the technologies developed in the research field and provide valuable research discussions. We first review BM-EL from the perspective of the technology-development process BM-EL methods into rule-based models, machine learning (ML) models, and deep learning

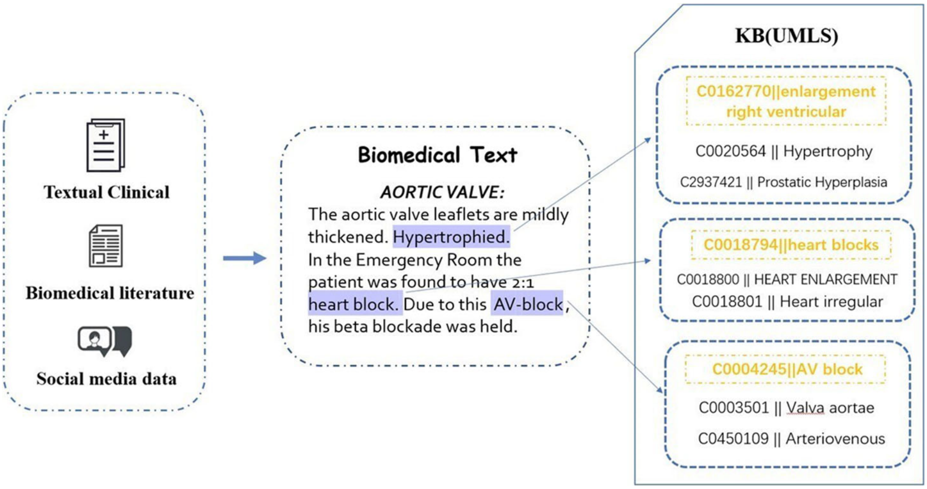

Textual Clinical
Biomedical literature
Social media data
Biomedical Text
AORTIC VALVE:
The aortic valve leaflets are mildly thickened. Hypertrophied.
In the Emergency Room the patient was found to have 2:1 heart block. Due to this AV-block, his beta blockade was held.
KB(UMLS)
C0162770||enlargement right ventricular
C0020564 || Hypertrophy
C2937421 || Prostatic Hyperplasia
C0018794||heart blocks
C0018800 || HEART ENLARGEMENT
C0018801 || Heart irregular
C0004245||AV block
C0003501 || Valva aortae
C0450109 || Arteriovenous

Figure 1 An illustration about the BM-EL task. Entity mentions detected from various types of medical texts are shaded; candidate entities for each entity mention in the knowledge base are indicated by blue dashed boxes on the right; their correct mapped entities are indicated by the bold yellow text

(DL) models. We primarily focus on the DL models because it has been flourishing over the past few years. Specifically, we conduct in-depth study of current BM-EL efforts and discuss four global modification and optimization methods for BM-EL model architectures: joint entity recognition and linking, graph-based global EL, cross-lingual architectures, and model-efficiency improvement. After that, we present a qualitative comparison among different categories of methods and analyze their advantages and disadvantages. Additionally, the biomedical domain involves various data categories, such as the corpora of the biomedical literature, social media medical texts, and disease and clinical records, which differ significantly from one another and lead to various challenges. This survey also focuses on the characteristics and challenges of these three categories of data, presents six of the most representative datasets that have been extensively used, and evaluates the results of different models on the corresponding datasets to achieve quantitative comparison.

Overall, our contributions can be summarized as follows:

1. We analyze the research conducted on BM-EL and summarize the issues and challenges of BM-EL. We further study BM-EL from the perspective of the technology-development process and present a comparison among different categories of methods and analyze their advantages and disadvantages.
2. We analyze the characteristics and challenges of datasets applied to different scenarios in biomedical and present six representative datasets. Afterward, we evaluate and compare the methods summarized in this paper on these datasets.
3. Based on exhaustive research and analysis of existing research work, we discuss the limitations of existing methods for BM-EL and investigate future research directions.

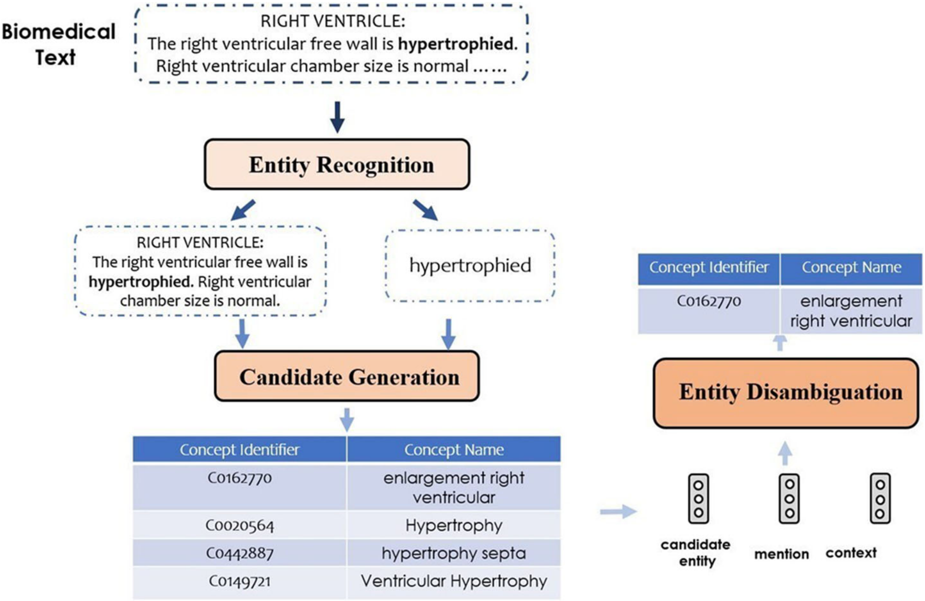
In this survey, we organize the overview as follows. We start with the definition of the BM-EL task in Section 1.2. In Section 1.3, we highlight the current problems and challenges of the BM-EL task. We then provide a systematic summary of the research progress related to BM-EL according to the technology-development process in Section 2 and summarize four methods for the overall revision and optimization of the BM-EL model in Section 3. In Section 4, we categorize commonly used BM-EL datasets into different types, introduce evaluation metrics, and present a quantitative analysis of different models' performances. Finally, we conclude the survey and suggest prominent directions for future work in BM-EL in Section 5.

1.2 Problem statement

EL is the task of mapping mentions in text documents to standard entities in a given knowledge base. A "mention" refers to language fragments that express entities in natural language text. An "entity" is a word or phrase that is clearly defined in the knowledge base and has a unique identifier. It can either be an ontology object in the real world (e.g., person or community) or an abstract concept (e.g., concept or definition). Then we present the BM-EL task in a formal definition as follows:

Definition 1 (Biomedical Entity Linking) Given a specified knowledge base (KB) in the biomedical field consisting of N entities $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$, a biomedical document $\mathcal{D}$ contains a set of recognized entity mentions $\mathcal{M} = \{m_1, m_2, \ldots, m_M\}$, and the task is to find the entity $e_i \in \mathcal{E}$ that $m_j \in \mathcal{M}$ refers to.

Typically, BM-EL has three steps. First is entity recognition, which identifies the corresponding biomedical mentions within the text. Second is candidate entity generation, which


Biomedical Text
RIGHT VENTRICLE:
The right ventricular free wall is **hypertrophied.**
Right ventricular chamber size is normal ……
Entity Recognition
RIGHT VENTRICLE:
The right ventricular free wall is **hypertrophied.** Right ventricular chamber size is normal.
hypertrophied
Candidate Generation
Entity Disambiguation
candidate entity
mention
context

Concept Identifier	Concept Name
C0162770	enlargement right ventricular
C0020564	Hypertrophy
C0442887	hypertrophy septa
C0149721	Ventricular Hypertrophy

Concept Identifier	Concept Name
C0162770	enlargement right ventricular

Figure 2 The architecture of BM-EL

generates a collection of relevant candidate entities for the mention from the knowledge base. Third is entity disambiguation, which ranks the candidate entities based on their relevance to the mention and selects the best candidate entity. We present a basic framework that applies to the majority of models for BM-EL in Figure 2. However, most papers involving EL, in the generic and biomedical fields, focus only on the entity-disambiguation task, and only a few studies focus on jointly performing entity recognition and disambiguation. Thus, in this survey, we primarily focus on how different techniques help improve candidate entity disambiguation.

1.3 Challenges in BM-EL

Although entity linking tasks have been investigated in the general domain and achieved great success, many challenges remain in the biomedical field. In practice, the semantic types and annotation criteria of biomedical datasets are often significantly different from the characteristics of data types in other domains. Consequently, directly replicating EL techniques that have performed well in other domains to study BM-EL problems may not achieve satisfying results. Accordingly, more challenges exist in the biomedical field as follows.

1. **Text Features:** Biomedical mentions and context usually have larger text spans than in the general domain. Moreover, more abbreviations, morphological variations, word order variations, and synonym variations exist. The highly complex terminological characteristic makes traditional EL techniques less effective in biomedical corpora [10].
2. **Entity Ambiguity:** Biomedical entity ambiguity refers to the fact that the same word or phrase can refer to different entities. In biomedical corpora, the same entity can have

several different names, making it even more challenging to normalize. It is difficult to link to the correct entity using only surface-level features, so many researchers try to use additional information, such as fine-grained types, to normalize entities [10].

3. **Corpora Limitations:** EL tasks usually require rich corpora to conduct experiments. For example, the extensively used Wikipedia corpora have no restrictions on the use of its annotated texts, and researchers do not need to consider privacy protection factors. However, the biomedical domain has fewer publicly available corpora and annotated data. For example, annotations of clinical records are often expensive and very restricted by privacy protection. Meanwhile, deep neural networks require a huge amount of training data to be effective, thereby posing a greater challenge.

Biomedical-related research involves various fields such as biology, chemistry, medicine, psychology, and statistics. Researchers consider a wide range of information, including genes, proteins, disorders, drugs, cells, body structures, and clinical information from EMRs[4]. The data characteristics of different biomedical fields significantly differ. The techniques used and the research focus are also completely different, so a huge amount of work is needed to summarize each one. For the reasons above, this paper focuses on the current state of EL and standardization research on biomedical literature, social-media medical texts, disease and clinical records, and other corpora in this field. In addition to the general challenges of the BM-EL task summarized in the previous section, we summarize the characteristics of the different types of medical text data and the corresponding technical difficulties.

Textual clinical data Electronic health records contain a description of the patient's medical history, family history, and symptoms, as well as the doctor's diagnosis based on the symptoms, physical and chemical indicators, etc. However, these data usually have a short context with much noise, e.g., misspelled words, incorrect grammar, abbreviations, and different variations of the same word. Moreover, for the same diagnosis and treatment plan, different physicians may record different results. Therefore, EL on clinical text is a current research focus.

Social-media-based medical data Social media and online health communities have a large amount of medical-related textual information, and people like to share various health experiences and consult related content online. For these data, we need to translate the social-media style text (e.g., “I feel my temperature is high” or “I feel like throwing up”) to formal medical style text (e.g., “fever” and “nausea”, respectively) [11]. Given the significant linguistic differences between medical terms and patient vocabulary [12] and the fact that social-media data often contain considerable noise, this task is very challenging [11] and has received much attention recently[11, 13–16].

Biomedical-literature data Researchers have published many highly valuable biomedical papers in a range of journals and magazines [17]. Biomedical literature is more standardized compared with clinical texts. Generally, it represents the technical research results that contain a large amount of specialized knowledge and unregistered terms. How to disambiguate terminology entities through EL is to be solved [18–21].

2 Biomedical entity linking

Reviewed the development of the models' paradigm, BM-EL technology is in constant progress, from the early rule-based models [22, 23] to the development models based on ML [18, 29, 30] . Ultimately, with the vigorous development of the DL technology in recent years, the DL model has become the most advanced in terms of performance in realizing the biomedical field as the most mainstream framework of EL method [15, 31, 40]. In summary, we categorize BM-EL methods into rule-based, ML, and DL models. This section focuses on each type of technical approach's essence and ideas and the different methods' core features. Table 1 summarizes the design choices BM-EL methods according to the proposed taxonomy.

2.1 Rule-based entity linking

Earlier was primarily implemented using string-matching or dictionary look-up approach to disambiguate entities. Table 2 shows a detailed comparison of different rule-based entity linking methods. While string-matching methods usually define templates based on setting spelling rules, word-formation rules, indicator words, prefix and suffix strings, followed by using templates for exact or partial matching. Dictionary-based methods use entries from existing dictionaries, which contain numerous vocabulary abbreviations, variants, synonyms etc. to identify and match entities. In this section, we review various rule-based approaches and summarize their models according to whether they are manual or automated in Table 2.

Different techniques can be used to assist both approaches (e.g., collecting concept mentions as additional synonyms from labeled data) [24, 44], and different string-matching techniques (e.g., string overlap and edit distance) [43]. MetaMap [45] and cTAKES [46] are the two most commonly used rule-based knowledge-intensive concept normalization tools that use rules to generate lexical variants for each noun phrase and then perform dictionary queries for each variant. Even for the rule-based approach, researchers have attempted to learn and generate the corresponding rules automatically. Islamaj Doan and Lu [47] used Lucene's search as a basis for disease-name normalization utilizing dictionary lookup and pattern matching. Kang et al. [23] focused on the obvious errors modeled by the dictionary query approach and designed an NLP model containing five rules. D'Souza and Ng [22] developed a multichannel filtering system based on the ShARe/CLEF dataset. The system defines 10 rules with various priorities to measure morphological similarity between mentions and entities for normalization purposes. Rule-based methods combined with ML techniques also exist, and most scholars use ML techniques to automatically resume criteria for selecting a suitable candidate. For example, Buyko et al. [25] transformed the gene-mention coordination problem into a sequence labeling task by using conditional random fields. Wermter et al. [48] developed a semantic similarity-scoring module in their Geno gene name-normalization system.

Although traditional rule-based methods achieve good results, most of them rely on a fixed, pre-defined approach, leaving them with the following problems. First, the rules are primarily artificially designed, and covering all regulations is impossible. Second, the design of the rules highly depends on the morphological characteristics of the entity, and distinguishing between morphologically similar but semantically different contexts is impossible. For example, the word "tender" originally means gentle and fragile, but in the sentence "Lymph nodes are enlarged but are not tender", it pertains to having hyperalgesia. Finally, entity rules in one field do not apply to another field, e.g., regulations designed on diseases do not apply to drugs.

Table 1 Summary of the design choices BM-EL methods according to our proposed taxonomy

Traditional method	Technology	Specific Category	Models
D’Souza and Ng [22]	Rule-based	String-matching	multi-pass sieve approach
Kang et al [23]	Rule-based	String-matching	multi-pass sieve approach
Leal et al9 [24]	Rule-based	Dictionary-based	CRF
Buyko et al. [25]	Rule-based,ML	Dictionary-based	CRF
Savov et al. [26]	ML	Classification	BOW
Stevenso [27]	ML	Classification	BOW
Gaudan et al. [28]	ML	Classification	SVM
Leaman et al. [18]	ML	Learning-to-rank	pLTR
Xu et al. [29]	ML	Learning-to-rank	RankSVM
Leaman et al. [30]	ML	Learning-to-rank	TaggerOne

Table 1 (continued)

DL-based Method	Embedding	Feature				Models
		SFS	SS	TS	CS	
Li et al. (2017) [19]	Word2VECs	√	√			CNN
Luo et al. (2018) [31]	Word2VECs	√	√			CNN
Schumacher(2020) [32]	ELMO	√				LSTM
Xu et al.(2020)[33]	BERT		√		√	–
Ji et al. (2019) [34]	BERT					
	BioBERT		√		√	–
	ClinicalBERT					
Zhao et al.(2019) [35]	Word2vec+	√				Bi-LSTM,CNN
	character-level					
Niu et al.(2019)) [36]	character-level	√	√			CNN
Pan et al. (2019) [37]	VSM	√	√			Rule-based,CNN
Murty et al. (2018) [38]	word embedding		√	√	√	MLP
Zhu et al. [10]	GloVe	√		√	√	Bi-LSTM
Ishani et al. [39]	word2vec	√			√	CNN
Tutubalin et al. [11]	HealthVec	√	√			LSTM,GRU
Attention						
Fakhraei et al. [40]	Siamese					
neural network	√	√			LSTM	
Angell et al. [41]	BERT		√		√	Graph-based
Vretinari et al. [42]	Graph-					
embedding	√	√		√	GraphSAGE	
R-GCN						
MAGNN						

Due to the limited space,"SFS" refers to the surface form similarity feature, "SS" refers to the semantic similarity feature,"TS" refers to the type similarity feature, and "CS" refers to the context similarity feature. "-" in the column of Models means the corresponding model uses a relatively simple algorithm to select the mapping entity, such as a linear combination of features

Table 2 Summary of rules-based entity linking methods

	Specific Category	Characteristic	Models	Manual / Automatic
D'Souza and Ng [22]	String-matching methods	Spelling rules, prefixes, word-formation rules, indicator words, and suffix strings	Multi-pass sieve approach (10 rules)	Manual
Kang et al. [23]			Multi-pass sieve approach (5 rules)	Manual
Islamaj Doan and Lu [47]				
Buyko et al.[25]			Lucene's search	Manual
Kate [43]			Levenshtein edit distance	Automatic
Buyko et al.[25]	Dictionary-based methods	Numerous vocabulary, abbreviations, variants, synonyms,nickname	CRF	Automatic
Leal et al[24]			CRF	Automatic

2.2 Machine learning based entity linking

Given the disadvantages of traditional rule-based methods, ML-based BM-EL tasks have been gradually developed and widely used. The ML-based BM-EL task primarily expresses entity mentions and candidate entities as feature vectors through manual design and statistical methods and then sorts and selects entities through various similarity-calculation methods. We can divide entity ranking techniques into two categories: classification and learning-to-rank.

The classification method converts the similarity-ranking problem into a classification problem [49] by using the trained classifier to mark the truth or falsity of entity referent-candidate entity pairs. When more than one referent entity pair is marked valid, then the referent-entity pair with the highest similarity is used as the disambiguation result by computing features such as bag-of-words and co-occurrence [26, 27]. For classification methods, a support vector machine (SVM) approach can be used to identify the separation hyperplane in the feature space that maximizes the interval to separate the positive and negative samples of the dataset. Gaudan et al. [28] used SVM trained on a "bag of words" model to resolve ambiguous global abbreviations and reported 98.5% accuracy. One disadvantage of all classification methods is that the output space tends to be small because the output space of a classification method must be the same as the number of concepts to be predicted. Moreover, the classification method tends to ignore the relationship between the candidate entity and the entity mention.

To avoid such problems, many systems utilize the learning-to-rank approach to rank the set of candidate entities. An ML-based approach, DNorm, was first proposed by Leaman et al. [18] for disease-word normalization. The basic idea of this model is to use pairwise learning to rank, which means comparing the similarity of mention found in the text to the entity concepts in the knowledge base and scoring them for ranking. Xu et al. [29] also designed a pairwise learning algorithm. They normalized each mention of each positive ADR to each entity in MedDRA by defining three features and by using RankSVM pairs. Leaman et al. [30] developed a generic named-entity identification and normalization toolkit, i.e., TaggerOne, based on a semi-Markov algorithm.

Although ML has made significant progress in performance and accuracy compared with rule-based methods, traditional ML methods have significant limitations. First, ML methods require complete and accurately labeled datasets, which are scarce and deadly for the biomedical field, especially in clinical medical texts or authoritative datasets in other languages such as Chinese. Second, ML methods are more dependent on specific types of domain knowledge, conferring difficulty in generalizing ML models trained this way to other types of biomedical domains.

2.3 Deep learning based entity linking

With the rapid development of DL techniques, neural networks are widely used because of their excellent generalization ability. These models have powerful feature-abstraction ability to learn practical and deeply distributed semantic information from texts, which makes DL excel in nearly all tasks [50]. Although DL-based BM-EL methods may differ in the implementation of technical details, their general steps are basically to generate various embeddings based on mention and candidate entities, followed by the use of the embeddings to calculate various features, and finally input the features into the DL models to obtain the optimum candidate entities [5]. Thus, we summarize the existing DL methods from three dimensions: embedding generation, feature extraction, and model architecture.

2.3.1 Embedding generation

Embedding encodes objects with low-dimensional numerical vectors and retains their meanings, which is convenient for upper-layer deep neural networks to process. DL-based methods usually need to capture semantic information and common knowledge such as lexical meaning and semantic roles from numerous unannotated corpora in advance by using various embedding techniques.

Recently, DL methods based on pretrained embeddings such as word2vec and ELMo have been effectively applied to many BM-EL tasks. Li et al. [19] and Luo et al. [31] pretrained word embeddings from large corpora with Word2VECs implementation. Miftahutdinov [51], Schumacher [32], and others used the ELMo to extend traditional word embeddings to contextual word embeddings and subsequently integrate it with existing task-specific architectures, thereby improving the state-of-the-art EL-architecture framework benchmarks.

All the above are traditional word embedding, which uses a one-way language model to learn language representation and uses only the preamble text information of a word to extract semantics. BERT pretraining model [52] is a multilayer two-way transformer encoder, which can complete sentence-level context word representation and has stronger semantic information-extraction ability. To improve the capabilities of NLP tasks in the biomedical domain, Lee et al.[53] pretrained large-scale biomedical texts and clinical notes and complete the BioBERT model based on the BERT model structure. Xu et al.[33] proposed an architecture that can consider morphological and semantic information, which consists of a candidate generator and a BERT-based list ranker. The BERT-based list-wise takes concept mentions and candidate concept names as input so it can process concepts that never appear in the training data. Ji et al. [34] regarded the BMEL task as a sentence-classification task and completed an EL architecture by fine tuning the BERT pretrained model. Wei et al. [54] proposed integrating BERT/BioBERT/ClinicalBERT models based on already fine tuned and trained BERT models into the corresponding local models and effectively using contextual semantic features.

Figure 3 shows typical architectures of word-level embedding and character-level embedding. As we can see from the figure, the word-level embedding methods cannot learn the character-structure features inside words. So many researchers have used character-level neural network methods to achieve character-level representation encoding by using the structural features inside words. Such methods are especially applicable for biomedical-concept standardization tasks because they can effectively solve the “out-of-vocabulary” (OOV) problem in noisy spoken medical texts. Zhao et al.[35] then used Bi-LSTM to stitch together pretrained word embeddings from Word2vec and character-level word representations from the convolutional neural network (CNN) to capture important character lexical information. Niu et al.[36] proposed a multitask character-level attention network to capture character-level features in OOV words and generate character-level attention weights on domain-related positions in text sequences to effectively improve the accuracy of normalization. Moreover, their existing EL methods combine entity-surface form, entity description, and entity-type information to learn entity embedding. Given that this part is strongly related to Section 2.3.2 feature introduction, we introduce it jointly in Section 2.3.2.

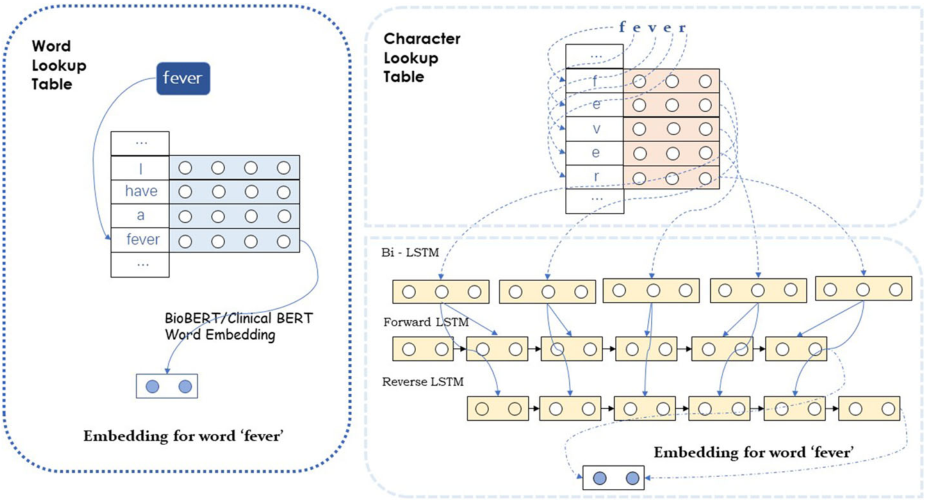


Figure 3 Typical architectures of word-level embedding and character-level embedding

2.3.2 Feature extraction

More and more methods are adopting different embedding techniques for extracting features to calculate the similarity between mentions and candidate entities, e.g., surface form, type, or synonym similarity, etc.

Learning through entities' surface forms is the most direct and easiest way to obtain entity embeddings. To solve the EL problem for small datasets, Pan et al. [37] proposed a CNN-based model with an ensemble of pretrained word vectors and two-step integration. The model enables a shallow structure and integration mechanism based on the perfect-matching morphological similarity approach to achieve proper linking in a limited training set.

A knowledge base contains entity instances, as well as entity types and hierarchies of types, which are popular features used by EL methods to learn entity embeddings. Murty et al. [38] investigated the use of hierarchy-aware loss functions on a deep neural network classifier to achieve integration of hierarchical type information in the embedding space of entities and types, thereby gaining statistical efficiency in predicting similar concepts and helping to classify rarer medical types. Zhu et al. [31] propose LATTE, a latent-type EL model, learning fine-grained types without direct supervision to assist in joint entity-disambiguation tasks. Vashishth [55] presents MEDTYPE, a fully modular system for pruning out overgenerated candidates in medical EL by predicting the semantic type of an entity mention.

Biomedical concepts may have multiple synonyms. Sung et al. [56] introduced BIOSYN, which utilizes the synonym-marginalization technique and the iterative candidate retrieval for learning biomedical entity representations, thereby maximizing the marginal likelihood of the synonyms present in top candidates. Yuan et al. [15] proposed to construct pretrained samples by collecting synonyms and definitions from knowledge base and sentence templates and then injecting them into generative language models. They proposed synonym-aware fine-tuning and decoding hints to improve the performance of the model by considering the similarity of the text.

2.3.3 Model architecture

Early DL-based techniques primarily utilize convolutional encoders. The main idea is to project mentions and entities as distributed vectors containing semantic information. After which, different backbone deep neural networks are applied to complete entity linking tasks. Limsopatham [13] applied pretrained word embeddings to the CNN and recurrent neural network (RNN) for the specification of medical concepts in social media texts and achieved optimum performance on multiple datasets. Li et al. [19] proposed a CNN-based architecture to compute the semantic similarity between candidate concepts and entity mentions and then rank the candidate entities generated by rule-based methods. Miftahutdinov and Tutubalina et al. [57] used a bidirectional RNN-based Encoder–Decoder model to implement the translation of the text on death certificates into medical codes. Luo et al. [31] proposed a multiview CNN model with a multitask shared structure to capture back and integrate valuable matching signals from different views, thereby solving the Chinese medical short-text normalization problem.

Afterward, LSTM has become the backbone model for many NLP applications and been widely used in BM-EL. Ishani et al. [39] designed a model based on triplet neural networks with a loss function that influences the relative distance constraint to identify positive and negative candidates concerning a disease mention. They explored the capability of in-domain sub-word-level information in solving the task of disease normalization. Tutubalin et al. [11] developed a direct task-specific end-to-end architecture for social-media medical-text normalization tasks. The architecture includes bidirectional long- and short-term memory, gated recursive units with attention mechanisms, and additional semantic similarity feature based on UMLS. Fakhraei et al. [40] used the LSTM model to map mentions and entities to a latent space while using the negative sampling technique to refine the embedding.

Most methods introduced above are based on similarity calculation, i.e., the model encodes mention and entity into the same vector space and calculates the similarity between

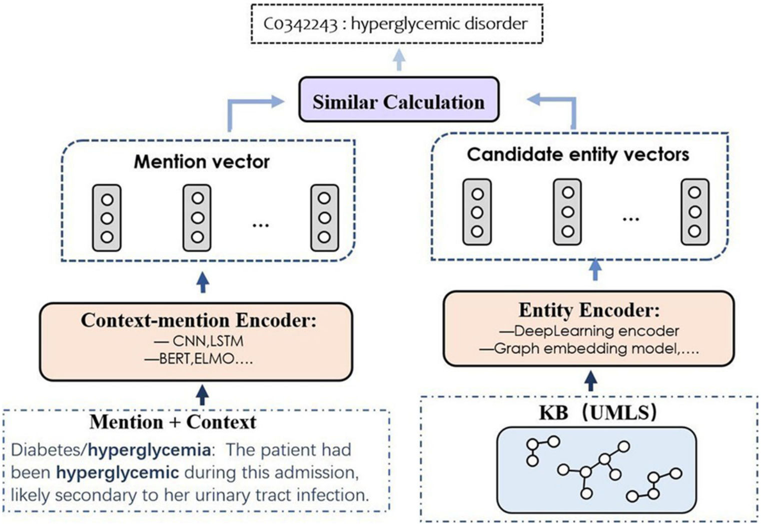

C0342243 : hyperglycemic disorder
Similar Calculation
Mention vector
...
Candidate entity vectors
...
Context-mention Encoder:
— CNN,LSTM
—BERT,ELMO....
Entity Encoder:
—DeepLearning encoder
—Graph embedding model,....
Mention + Context
Diabetes/hyperglycemia: The patient had been hyperglycemic during this admission, likely secondary to her urinary tract infection.
KB （UMLS）

Figure 4 Generalized candidate entity ranking neural architecture

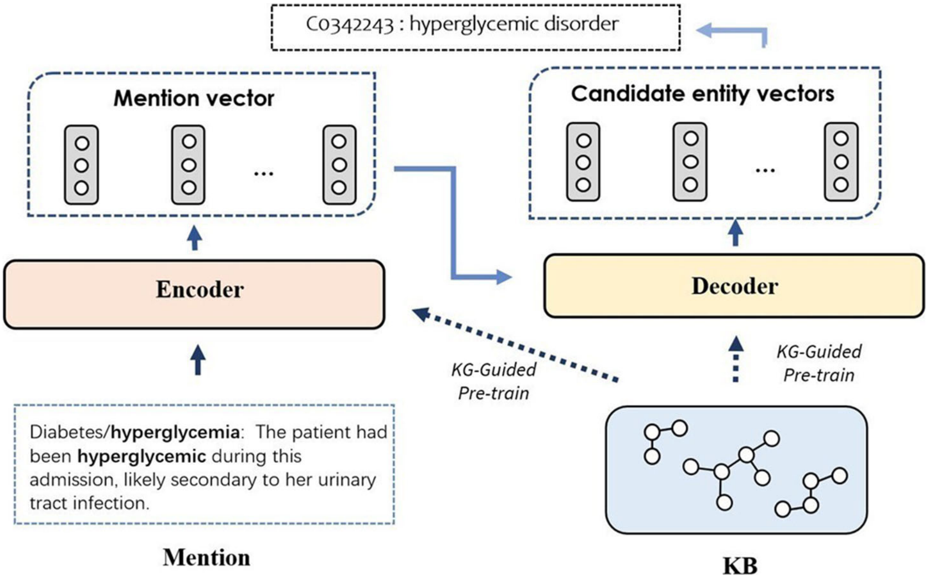

C0342243 : hyperglycemic disorder
Mention vector
Candidate entity vectors
Encoder
Decoder
KG-Guided Pre-train
KG-Guided Pre-train
Diabetes/**hyperglycemia**: The patient had been **hyperglycemic** during this admission, likely secondary to her urinary tract infection.
Mention
KB

Figure 5 Generative EL architecture

the embedded representations. Figure 4 depicts the typical architecture of the ranking component. Such methods need to consider the problem of negative sampling in the training phase and consume substantial storage in the inference phase.

Different from the ideas of the above methods, applying generative methods to EL tasks in the biomedical field has gained considerable attention. Researchers regard the EL task as a Natural Language Generation task (NLG), where a piece of text containing mention is input, and the linked entity name is output. Figure 5 depicts a typical architecture for generative EL methods. Cao et al. [58] proposed GENRE, a seq2seq EL method in the general domain pretrained on the Wikipedia EL dataset. However, using such a method directly in biomedical domain has many limitations due to the lack of a massive expert-annotated dataset and the high number of entity synonyms. Yuan et al. [15] used the Encoder-Decoder structure of Transformer to accomplish biomedical seq2seq EL. The model reduces the need for corpus size in the training phase by composing input sequences with synonyms and explanations collected from the knowledge base and utterance templates. The BioBART model proposed by Yuan et al. [16] also applies a generative approach. The model is based on the architecture of BART, a classic model for NLG tasks, and removes some of the pretraining tasks considering the difference in model-performance requirements for different tasks. The model is pre-trained on the PubMed corpus and reaches a new SOTA on several biomedical entity-linking datasets. However, all generative EL methods have a common feature the training process requires vast computational resources to achieve competitive performance.

3 Architecture modifications and optimizations

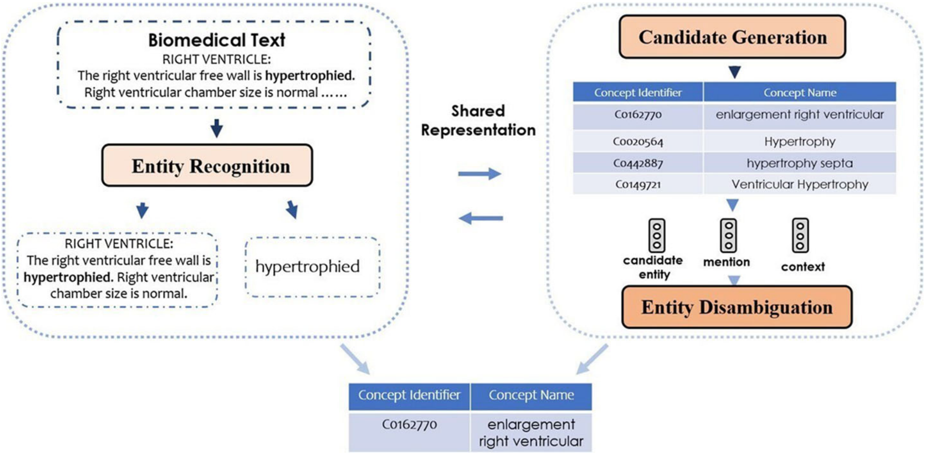
In the previous section, we introduce the main techniques according to the lineage of technology development. In this section, we also focus on the current important methods of overall model revision and technology optimization in the BM-EL field: joint entity recognition and linking, graph-based global EL, cross-lingual architectures, and model-efficiency improvement.

3.1 Joint entity recognition and disambiguation architecture

Existing research treats entity recognition and disambiguation as separate steps in EL. This decoupling approach uses pipeline models to implement entity disambiguation and entity recognition separately, leading to error cascades and a lack of mutual benefits.

State-of-the-art studies have shown that the joint modeling of biomedical-named entity recognition and disambiguation has advantages over pipeline implementation owing to their mutually beneficial relationship. Figure 6 depicts a typical architecture of joint entity identification and disambiguation architecture. Leaman and Lu [30] used a joint scoring function for biomedical-named entity recognition and disambiguation. Lou et al.[31] proposed a transformation-based model that treats the output construction process as an incremental transformation process. However, both methods rely heavily on manual features and use simplistic methods to achieve union, which cannot encode complex features. Zhao et al.[35] proposed a novel deep neural multitask learning framework based on CNN, LSTM, and Word2vec methods to provide necessary support between biomedical-named entity recognition and disambiguation joint.

Recently, specific models proposed the use of multitask analysis of recursive inference to achieve overall optimization of the model. Rajani et al.[59] argued that integration techniques may be superior to component techniques, using stacks to store multiple model systems' output and additional overlay features to evaluate the system output and train a meta-classifier. The CNN-based method proposed by Niu et al. [36] combines the averaging module of the word vector representation of a mention and the maximum pooling module after CNN feature extraction with the maximum pooling operation of the sentence representation where the mention is located. The combined result is then jointly fed into two multilayer perceptrons to obtain the output. Wiatrak [20] and Mrini K [60] treated EL as a multitask model, with the difference that Wiatrak [20] proposed a new task of mention detection and entity typing. The mention detection is performed first, and then a multitask


Biomedical Text
RIGHT VENTRICLE:
The right ventricular free wall is **hypertrophied**.
Right ventricular chamber size is normal ……
Entity Recognition
RIGHT VENTRICLE:
The right ventricular free wall is **hypertrophied**. Right ventricular chamber size is normal.
hypertrophied
Shared Representation
Candidate Generation
candidate entity
mention
context
Entity Disambiguation

Concept Identifier	Concept Name
C0162770	enlargement right ventricular
C0020564	Hypertrophy
C0442887	hypertrophy septa
C0149721	Ventricular Hypertrophy

Concept Identifier	Concept Name
C0162770	enlargement right ventricular

Figure 6 Joint entity recognition and disambiguation architecture

Table 3 Summary and comparison of joint methods

	Main Task	Auxiliary Task	Models	Data type
Leaman and Lu [30]	Named Entity Recognition	Named Entity Recognition	Semi-Markov structured	Not specific
Rajani et al.[59]	Stacking multiple EL system	–	Stacking with Auxiliary Features	EMR data
Niu et al. [36]	Named Entity Recognition	Character-level weights learning	Multi-Task Attentional	Social-Media-Based Medical Data
Lou et al.[31]	Disease Disambiguation	Procedure Disambiguation	Multi-view CNN	Disease and Procedure data
Zhao et al.[35]	Named Entity Recognition	Named Entity Recognition	Deep neural multi-task Learning framework	Disease data
Wiatrak [20]	EL	Entity Typing (ET) Mention Recognition (MR)	Hierarchical Multi-task model	Biomedical Literature data
Mrini K et al. [60]	EL Match Prediction	Mention Detection Re-ranking approach	Multi-task learning Medical Data	Social-Media-Based

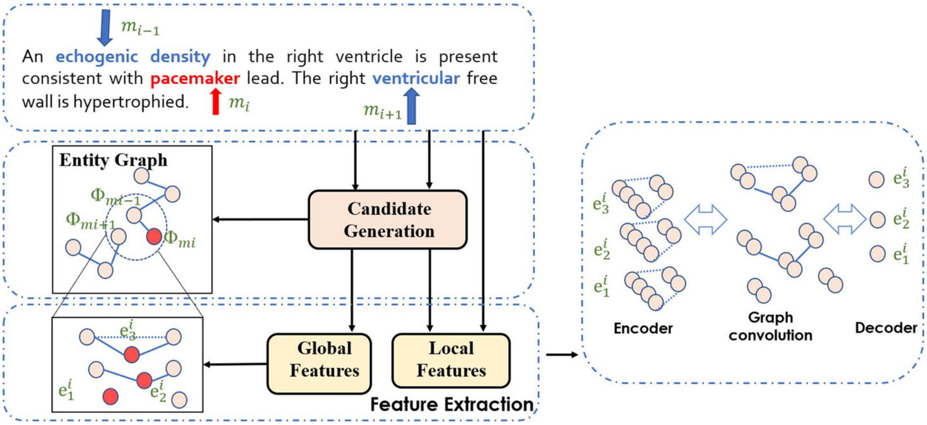
framework is trained using weighted targets, followed by reranking by entity typing. Meanwhile, Mrini K et al. [60] used the multitask learning of a hierarchical task (entity prediction, entity classification, and entity disambiguation in a three-feature progressive representation) to implement it jointly. Table 3 shows detailed comparison of different methods summarized in this section, including main tasks, auxiliary tasks, model methods, and data types.

3.2 Graph-based global architectures

Conventionally, EL methods can be categorized into local and global models. Local methods primarily obtain the information contained in the entity mention and its surrounding words in the specified window. Each mention is independent of the other, and the information is not interactive [61–64]. The local model focuses only on how to link the entities extracted from the text to the knowledge base, ignoring the semantic connections between different entities located in the same document. Conversely, global models encourage all alleged target entities in a document to be thematically consistent and disambiguate by computing the thematic consistency, entity relatedness, transfer probability, and entity popularity features between different target entities [65–70].

Global models usually build entity graphs based on a knowledge base to capture all identified allegedly coherent entities in a document. Figure 7 shows the framework of graph-based global model. The nodes in the graph represent entities, and the edges represent their relationships. Generally, based on the document and its entity graph, we extract local and global features for each candidate entity, and the association correlations among entity mentions, candidate entities, entity mentions, and candidate entities are used for collaborative inferences. Then the obtained features are encoded, and graph convolution and other methods are used to conduct graph-based ranking algorithm to select the candidate entities with the highest score. In the following section, we will introduce graph-based global architectures used by BM-EL.

Current supervised methods require a mass of manually labeled training data, which is challenging for medical data. To address this problem, Jin et al. [71] proposed a new unsupervised collective inference method, which works by obtaining well-structured ontologies



An echogenic density in the right ventricle is present consistent with pacemaker lead. The right ventricular free wall is hypertrophied.

Figure 7 Framework of graph-based global model. The inputs of a set of mentions in a document are listed at the top. The words in red indicate the current mention m_i where m_{i-1}, m_{i+1} are neighbor mentions, and $\Phi\left(m_i\right) = \left\{e^i_1, e^i_2, e^i_3\right\}$ denotes the candidate entity set for m_i

in the class of hierarchical and relational structures, as well as good semantic relationships among ontologies. The knowledge base graph is constructed after the relationships and subjected to similarity and entity ranking.

Angell et al. [41] proposed a new process based on clustering-based inference to implement joint EL, which is primarily based on an inference approach to construct a graph where the union of mentions and entities is constructed as nodes in the graph. The edges in the graph indicate the correlation weights between nodes, and multiple similar nodes are gathered into a combination by a clustering approach. As long as one mention in the aggregated mention group is linked to the correct entity, the whole cluster can be correctly classified, and the zero-shot EL task can be effectively solved by such joint link-prediction methods.

Graph-representation learning has shown promising results in various representation learning tasks on knowledge bases. D. Pujary et al. [72] utilize the graphical structure of MeSHR and taxonomy by using state-of-the-art neural graph embeddings (GCN and node2vec) to represent disease names. The neural-named entity recognition is now combined with graph-based EL methods through multitask learning to improve the disease-name normalization problem. Vretinari et al. [42] on top of GraphSAGE [73], R-GCN [74], and MAGNN [21] to build the new ED-GNN model. The model represents the entities mentioned in the text fragments as query graphs and learns how to generate node embeddings by aggregating rich structural and semantic information from the neighboring regions of each node. The model is a robust spatially invariant aggregation function, and an effective negative sampling strategy is designed to identify negative sampling and improve the disambiguation capability of the model.

3.3 Cross-lingual architectures

Currently, the vast majority of entities in the BM-EL tasks are available only in English. For example, about 70% of terms in the UMLS[75, 76] meta thesaurus are from English, about 11% are from Spanish, and less than 3% are from other included languages. The cross-lingual EL approach [77] is valuable in advancing the field of BM-EL by using supervised signals from multiple languages to train models in the target language. Early CLEF challenges include non-English biomedical-text normalization tasks. However, in 2015 and 2016, most teams rely on tools such as Google Translate and Bing TranslatorDENG [78–80], which help achieve good results for the tasks but have more limitations, such as web translators that cannot handle clinical documents or strictly privacy-protected data that cannot be processed online.

To overcome the limitations of translation tools, many neural network models have been used for the task of cross-lingual EL in the biomedical field. Roland Roller et al. [81] proposed a conceptually normalized cross-language candidate search based on a character-based neural translation model trained on multilingual biomedical terminologies. The model is trained using UMLS's Spanish, French, Dutch, and German versions. Fangyu et al. [82] introduced a novel cross-lingual biomedical entity task (XL-BEL) to establish a benchmark for cross-lingual entity representations in biomedical domains in 10 languages with broad coverage and reliable evaluation of current SotA on XL-BEL biomedical-entity representations on XL-BEL. An effective transfer-learning scheme is also proposed, which uses translations of generic domains to improve the cross-linguistic capabilities of domain-specialized representation models. Borchert et al. [83] addressed the problem of a very sparse corpus annotated at the entity-mention level and its mapping to concepts in languages other than English by proposing a hybrid EL system, which is based on the NER pipeline of standard transformers for mention detection. Accordingly, two complementary candidate-generation methods are provided: a TF-IDF vectorizer based on character n-grams and a cross-lingual SapBERT model. Finally, using a rule-based reranker, a translation task for Spanish language clinical-pathology reports is achieved.

3.4 Improvements on model efficiency

Many BERT-based models have been used to improve BM-EL and achieve state-of-the-art results on many BM-EL datasets. However, these models are computationally expensive, and the improvements made by fine tuning bring high computational costs and memory usage. Thus, aside from model-accuracy improvements, researchers are also interested in optimizing the number of model parameters and improving model inference speed. In Table 4, we summarize models' backbone architectures, the number of parameters and compare their performance in terms of accuracy and model inference time in this section.

Lai et al. [84] conducted prior experiments disrupting word order and limiting the range in attention mechanism. After which they found that the results of BERT-based models are not significantly degraded. Accordingly, they concluded that for BM-EL tasks, the rich syntactic and semantic information brought by the vast number of parameters used in the BERT-based model are not fully used. Thus, their proposed model uses ResCNN as the backbone network to encode the embedding representation, which is initialized by PubMedBERT [85]. Experimental results show that the model achieves similar performance to the BERT-based model while the number of parameters is reduced to 1% and the inference speed increases by up to 21.3 times.

Chen et al. [86] also used CNN as the backbone network of the model to significantly reduce the model parameters. They further introduced more features into the embedding representation of mention and entity to improve the model performance. By designing the Alignment Layer in the Ranking stage, the model computes the attention on each token of mention and entity names separately, thereby solving the problems such as having multiple names for the same entity (including interference from different word orders) in the biomedical domain. The number of parameters is significantly optimized, whereas the performance remains similar to that of BERT-based models. Experimental results show that the number of model parameters is reduced by up to 99%, and the inference time is improved by up to 12.3 times.

Table 4 Efficiency comparison among models' number of parameters, accuracy and inference time on different datasets

Model	Category	Parameters	NCBI-d			BC5CDR		MedMentions	
			Acc	Time		Acc	Time	Acc	Time
				CPU	GPU		GPU		GPU
BERT(base)	BERT based model	110M	88.7	443s	83s	–	–	–	–
SAPBERT	BERT based model	110M	92.3	534s	58s	95.0	342s	50.4	6269s
Bhowmik et al.	BERT based model	110M	–	–	–	80.7	72s	68.4	1832s
Chen et al.	CNN	4.6M	89.6	38s	22s	–	–	–	–
Lai et al.	CNN init. w/ PubMedBERT	1.7M	92.4	33s	18s	95.1	90s	53.5	1565s

Models are grouped by their backbone architectures

Bhowmik et al. [87] proposed using a dual encoder to perform candidate-entity generation and entity disambiguation through a single model in an end-to-end EL paradigm. The model completes EL with only one document traversal, and the training process is three times more efficient than other models with the same amount of parameters. Additionally, the model can compare mention with all entities in the knowledge base during the inference phase, thereby saving time in candidate-entity generation and improving the inference speed by up to 25 times compared with the baseline approach.

Ye et al. [88] consider that Pre-trained Language Models cannot well recall rich factual knowledge of entities exhibited in large-scale corpora, especially those rare entities. They build a simple but effective Pluggable Entity Lookup Table (PELT) by aggregating the entity's output representations of multiple occurrences in the corpora. This method merely consumes 0.2% – 5% pre-computation compared with previous works, and it also supports the vocabulary from biomedical publication.

4 Datasets and performance evaluation

In Section 1.3, we introduce the characteristics of three datasets commonly used in the biomedical field, including the corpora of biomedical literature, social-media medical texts, and disease and clinical records. They differ significantly from one another and lead to different challenges. In this section, we introduce six representative datasets and evaluate the results of different models on related datasets.

4.1 Datasets

According to different corpus sources, we classify the datasets commonly used for BM-EL tasks into the following types: biomedical literature, EMRs, social media datasets, etc. These corpora have various text features. For example, literature has more complex proper nouns, more abbreviations in electronic medical records, and more colloquial expressions in social media data. We summarize the statistical information of different datasets in Table 5 and introduce their details as follows.

MedMentions dataset Constructed by Mohan and Li [76]. It is one of the largest BM-EL datasets available, and includes 4392 English abstracts from PubMed with contains 352,496 mentions. Each mention is linked to a unique entity in the UMLS knowledge base. Researchers usually use the St21pv subset, including fewer mentions, CUIs, and a total of 21 semantic types of entities. For the partitioning of the dataset, researchers follow the official 60%/20%/20% ratio to obtain the train/dev/test set.

Biocreative V CDR dataset Constructed by Li et al. [89]and is widely used in Named Entity Recognition and EL tasks. It is a corpus of the biomedical literature derived from 1500 English language articles in PubMed, containing 4409 annotated chemicals and 5818 annotated disease entities. All the mentions in the dataset are linked to MeSH (a subset of UMLS). The articles are equally distributed into train/dev/test sets.

NCBI disease corpus Constructed by Dogan et al. [90] It is an extensively used entity-linking dataset obtained from biomedical literature. The dataset contains 793 abstracts of the biomedical literature, where each mention is linked to MEDIC ontology [91]. Notably,

Table 5 Statistical information of different BM-EL datasets

Dataset	Corpus Type	Year	Documents	# Mentions	Unique Entities	KB	# Entities
MedMentions	PubMed Abstract	2019	4,392	352,496/203,282[a]	34,724/25,419[a]	UMLS 2017	3,271,124
BC5CDR	PubMed Article	2016	1,500	15,935/12,850[b]	9,149	MeSH[c]	4,409/5,818[b]
NCBI	PubMed Abstract	2014	793	6,892	790	MEDIC[d]	11,915
ShARE/CLEF	EMR	2013	298	11,167[e]	–	SNOMED-CT	88,140
COMETA	Social media	2020	–	19,911	7,648	SNOMED-CT	–
AskAPatient	Social media	2016	–	8,662[f]	–	SNOMED-CT/AMT	1,036

[a] The form is FULL set/ST21pv subset

[b] The form is BC5CDR-Chemical / BC5CDR-Disease

[c] MeSH is a subset of Unified Medical Language System

[d] MEDIC is a combination of Mesh descriptor and OMIM identifier

[e] About 30% are NIL(unlinkable)

[f] It refers to total phrases in the dataset

Bold entries represent the best-performing models in the same category in the NCBI dataset, each annotation mention is linked to an entity in the knowledge base. A typical dataset split division is 593/100/100 [56].

ShARe/CLEF eHealth Challenge corpus Constructed by Pradhan et al. [92]. Unlike the datasets mentioned above, ShARe/CLEF is an electronic medical-record dataset containing 298 clinical reports. In each report, the disorder mention is linked to the corresponding entity in the SNOMED-CT knowledge base. If no corresponding entity exists, it is labeled as “CUI-less” (about 28.2% in the training set and 32.7% in the test set [86]). The dataset is separated into training (199) and test (99) subsets [22].

COMETA and AskAPatient COMETA [93] and AskAPatient [13] are datasets obtained from social media and forums. More types of entity-linking datasets in the biomedical domain are attracting researchers’ interest. Unlike English literature and electronic medical-record datasets, as the general public does not have a rich professional background, their descriptions are relatively vague and imprecise. Their language styles are also relatively uncritical, posing new challenges for BM-EL models.

4.2 Evaluation metrics

The most intuitive way to evaluate an EL model is the proportion of mentions correctly linked to the corresponding entity, which is the Precision (P). We also need to consider the performance of candidate-entity generation, so we introduce the metric of entity Recall (R). We consider these two metrics together and calculate the F1 value to evaluate the model’s overall performance.

We denote $\mathbf{M}$ as the set of all mentions, $\mathbf{M}_c$ as the set of mentions correctly linked to the entity, ε as all generated candidate entity sets, ε_c as candidate entity sets containing the correct entity. We define the above-mentioned metrics formally as follows.

$$Precision = \frac{\mid \mathbf{M}_c \mid}{\mid \mathbf{M} \mid} \tag{1}$$

$$Recall = \frac{\mid \varepsilon_c \mid}{\mid \varepsilon \mid} \tag{2}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

where $\mid \mathbf{M} \mid$ represents the item number of a set $\mathbf{M}$, vice versa.

For models that consider named-entity recognition and entity disambiguation, the F1 metric provides an excellent overall picture of the model’s performance. Researchers usually use the Accuracy metric for models that consider only entity disambiguation. It is calculated in the same way as Precision.

$$Accuracy = \frac{\mid \mathbf{M}_c \mid}{\mid \mathbf{M} \mid} \tag{4}$$

4.3 Performance analysis

We present a series of representative BM-EL methods and collect experimental results from the original papers in Section 4.1, and summarized in Table 6.

As shown in Table 6, the SOTA performance is found to be achieved on almost all datasets for the DL-based approach compared with the rule-based and ML based methods, which proves that the DL technique works very well for the medical EL task. Specifically,

Table 6 Reported results for BM-EL evaluation

	ShARe/CLEF		BC5CDR		NCBI		MedMentions[a]	
	F1	Acc	F1	Acc	F1	Acc	Acc	MRR
Leaman et al. [18]	–	–	–	–	0.782	–	–	–
Li et al. [89]	0.903	–	–	–	–	0.861	–	–
Li et al. [19]	0.903	–	–	–	–	0.861	–	–
Liu et al. [82]	–	–	–	0.77	–	–	–	–
Murty et al. [38]	–	–	–	–	–	–	**–/0.754**[b]	–/0.597
Luo et al. [31]	0.898	–	–	–	–	–	–	–
Zhao et al. [35]	–	–	**0.892**	–	0.882	–	–	–
Mondal et al. [39]	–	–	–	–	–	0.900	–	–
Wright et al. [94]	–	–	0.834	0.88	0.84	0.878	–	–
Chen et al. [86]	–	–	–	–	–	0.898	–	–
Schumacher et al. [32]	0.700	0.78	–	–	–	–	–	–
Sung et al. [56]	–	–	–	0.93	–	0.911	–	–
Angell et al. [41]	–	–	–	–	–	–	–/0.487	**–/0.781**
Vashishth et al. [55]	–	–	0.026[c]	–	0.874	–	0.100[c]	–
Wiatrak et al. [20]	–	–	0.639	0.92	–	–	0.443/0.415	0.682/0.764
Vretinaris et al. [42]	0.825	**0.88**	0.874	–	0.874	**0.924**	–	–
Varma et al. [95]	–	–	–	0.92	–	–	–	–/0.748
Yuan et al. [15]	–	–	–	–	**0.891**	–	–	–
Lai et al. [84]	–	–	–	**0.95**[c]	–	**0.924**	–	–
Chen et al. [86]	**0.904**	–	–	–	–	0.896	–	–
Bhowmik et al. [87]	–	–	0.752	–	–	–	0.564/–	–
Abdurxit et al. [96]	–	–	–	–	–	0.913	–	–

[a] The form is FULL set/Subset

[b] The dataset here is a subset of a pre-release MedMentions version, which is different from the other models

[c] Not available in the original paper. The result is calculated by the author manually through other reported results

Bold entries represent the best-performing models in the same category

Vretinaris et al. [42] achieves the best results on the ShARe/CLEF and NCBI datasets, and owns the leadership on the BC5CDR dataset, proving that applying graph neural networks can collectively learn the contextual information and structural interdependence across mentions and also capture the unique and informative contextual information of entities in a medical knowledge base. Zhao et al. [35] achieves the best results on the BC5CDR dataset and owns the leadership on the NCBI dataset, showing that the joint modeling of medical named-entity recognition and normalization has advantages over pipeline implementation owing to their mutually beneficial relationship. Owing to the selection of the backbone network and the introduction of more high-quality features, Chen et al. [86], Lai et al. [84] maintain a high level of model performance despite the significant improvement in model inference speed and reduction in model size, which can bring comparative advantages to the system deployment and application.

As shown in Table 7, the generative approach significantly outperforms other methods in terms of recall, and Yuan et al. [15] established the most muscular performing model so far.

Table 7 Reported Recall@1 results comparison between Generative EL and Traditional EL methods

	MedMentions	BC5CDR	NCBI
Leaman et al. [18]	–	–	0.763
Wright et al. [94]	–	0.805	0.818
Angell et al. [41]	0.475	–	–
Vashishth et al. [55]	0.053	0.013	0.065
Wiatrak et al. [20]	0.4237	0.6291	–
Bhowmik et al. [87]	0.564	0.744	
Vretinaris et al. [42]	–	0.881	0.889
Yuan et al. [15]	–	**0.933**	**0.919**
Yuan et al. [16]	**0.7178**	0.9326	0.8990

Bold entries represent the best-performing models in the same category

This model indicates that the self-regression-based BM-EL method can generate candidate entities more effectively, and the combination of synonym-aware fine tuning can accomplish entity disambiguation well. However, seq2seq EL models require heavier computational resources during training. The generative approach is still in the exploration stage in BM-EL tasks. It is a valuable research direction to reduce the model size and computational-resource requirements while maintaining the strong performance of generative models.

Furthermore, no optimum EL systems can perform well on all datasets owing to the varied characteristics of different biomedical datasets, such as clinical document length and a number of entities mentioned per document. For a given dataset linked by entities, we should use suitable models to obtain advanced results according to the data characteristics.

5 Conclusion and future directions

In this survey, we summarize and review the recently proposed BM-EL models. We first discuss BM-EL from the perspective of technology development and technology path. Then, we systematically review the current representative BM-EL methods in each category above. Lastly, we summarize commonly used BM-EL datasets by different source corpora types, compare the reported results of the different models under various datasets, and present results analysis. We believe that numerous barriers need to be overcome, and much space for future improvement exists. Below, we summarize the limitations and shortcomings of existing BM-EL methods and discuss promising future research directions.

1. **Weak supervision/NO supervision EL:** Traditional supervised models require a large amount of already labeled data for training, which is costly for clinical medical data. To tackle this problem, weak supervision [29, 30] is a potential approach, which is an intermediate learning approach between supervised and unsupervised that uses heuristics, knowledge base, crowdsourcing, and other sources to automatically create labeled training data to reduce the burden and cost of labeling training data. Dong et al. [97] proposed an ontology-based and weakly supervised approach for rare-disease phenotyping from clinical notes and achieved the optimum result in extracting rare-disease UMLS phenotypes from MIMIC-III discharge summaries. This finding may suggest that a study can be conducted to learn how to complete tasks using small sample data or weakly supervised data using weakly supervised or unsupervised methods. Du et al. [98] consider supervision of entity mention's boundary as unnecessary for applications

like semantic search engines and chatbots, which only utilize the information from the set of entities. Their results indicate that mention-aware models and mention-agnostic models achieve comparable performance under a new-performed evaluation formulation. In biomedical domain, several applications focus less on mention-span, and extract semantic information mainly from the set of entities. Adapting this method might be helpful for these downstream tasks, and could be a perspective of future work.

2. **Multisource heterogeneous text data:** Biomedical text corpora originate from multiple data sources and contain diverse structured and unstructured forms of data. Currently, most BM-EL systems focus on how to detect entity mentions from unstructured documents and map them to the knowledge base, e.g., biomedical literature or EMRs. However, biomedical data also has data types such as tables and lists in the hospital's electronic health system. Different types of data have different characteristics, some detailed data such as tables are semistructured texts with almost no textual context. Only a few people have focused on the task of BM-EL with structured or semistructured data. Therefore, it is an interesting direction of future work to develop specific techniques to handle connected entities.
3. **Model robustness and efficiency:** Improving the EL model's robustness and efficiency is receiving significant attention. For BM-EL, robustness means achieving consistent performance over various datasets. For example, consistency can be maintained across different textual structures, such as social-media data, clinical texts, or medical literature. Furthermore, most works on BM-EL lack an analysis of computational complexity. However, efficiency and scalability are essential for real-time and large-scale applications. Although Lai et al. [84] and Chen et al. [86] aimed to determine how to improve the efficiency of EL tasks, they did not perform tests on large datasets. Therefore, an essential direction for future research is to investigate the design of systems that substantially improve the efficiency and scalability of BM-EL systems while maintaining high accuracy and precision. Ayoola et al. [99] propose an efficient end-to-end model, completing Named Entity Recognition, Entity Typing and Entity Disambiguation in only one pass on input documents on large-scale KGs like Wikidata. Their model is 60x faster than previous systems. In addition to aforementioned research perspectives, Dong et al. [100] propose a novel method using cache embedding table to reduce communication cost during distributed knowledge graph embedding training process. Adapting these methods to biomedical domain, especially on large-scale or distributed Biomedical KGs, might be helpful on solving the problem of high computational costs.

We hope that this survey demonstrates the current status and limitations of existing BM-EL research and provide insight for researches to conduct in-depth research in this area.

Author Contributions Jiyun Shi: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing for Original Draft. Zhimeng Yuan: Methodology, Validation, Formal analysis, Visualization. Wenxuan Guo: Methodology, Validation, Formal analysis. Meihui Zhang: Review and editing, Project administration. Chen Ma: Review and editing. Jiehao Chen: prepared the table, Investigation, Data Curation.

Funding No Funding.

Availability of supporting Data Not applicable.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for Publication Not applicable.

Competing interests We declare that authors have no known competing interests or personal relationships that might be perceived to influence the discussion reported in this paper.

References

1. Reddy, C.K., Aggarwal, C.C.: Healthcare data analytics (2015)
2. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Res. **32**(suppl_1), 267–270 (2004)
3. Huang, M.-S., Lai, P.-T., Li, P.-Y., You, Y.-T., Tsai, R.T.-H., Hsu, W.-L.: Biomedical named entity recognition and linking datasets: survey and our recent development. Brief. Bioinform. **21**(6), 2219–2238 (2020)
4. Tsai, R.T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsiang, J., Sung, T.-Y., Hsu, W.-L.: Various criteria in the evaluation of biomedical named entity recognition. BMC Bioinformatics **7**(1), 1–8 (2006)
5. Shen, W., Li, Y., Liu, Y., Han, J., Wang, J.: Yuan, X. Entity linking meets deep learning, Techniques and Solutions (2021)
6. Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., Biemann, C.: Neural entity linking: a survey of models based on deep learning arXiv e-prints (2020)
7. Rao, D., Mcnamee, P., Dredze, M.: Entity linking: finding extracted entities in a knowledge base.Springer Berlin Heidelberg (2013)
8. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: issues, techniques, and solutions. IEEE Trans. Knowl. Data Eng. **27**(2), 443–460 (2015)
9. Al-Moslmi, T., Ocaa, M.G., Opdahl, A.L., Veres, C.: Named entity extraction for knowledge graphs: a literature overview. IEEE Access **8**(1), 32862–32881 (2020)
10. Zhu, M., Celikkaya, B., Bhatia, P., Reddy, C.K.: Latte: latent type modeling for biomedical entity linking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9757–9764 (2020)
11. Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., Malykh, V.: Medical concept normalization in social media posts with recurrent neural networks. J. Biomed. Inform. **84**, 93–102 (2018)
12. Karimi, S., Wang, C., Metke-Jimenez, A., Gaire, R., Paris, C.: Text and data mining techniques in adverse drug reaction detection. ACM Computing Surveys (CSUR) **47**(4), 1–39 (2015)
13. Limsopatham, N., Collier, N.: Normalising medical concepts in social media texts by learning semantic representation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers), pp. 1014–1023 (2016)
14. Miftahutdinov, Z., Tutubalina, E.: Deep neural models for medical concept normalization in user-generated texts (2019)
15. Yuan, H., Yuan, Z., Yu, S.: Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. arXiv (2022)
16. Yuan, H., Yuan, Z., Gan, R., Zhang, J., Xie, Y., Yu, S.: BioBART: pretraining and evaluation of a biomedical generative language model. arXiv (2022)
17. Almeida, T., Antunes, R., F Silva, J., Almeida, J.R., Matos, S.: Chemical identification and indexing in pubmed full-text articles using deep learning and heuristics. Database **2**022 (2022)
18. Leaman, R., Islamaj Doğan, R., Lu, Z.: Dnorm: disease name normalization with pairwise learning to rank. Bioinformatics **29**(22), 2909–2917 (2013)
19. Li, H., Chen, Q., Tang, B., Wang, X., Xu, H., Wang, B., Huang, D.: Cnn-based ranking for biomedical entity normalization. BMC Bioinformatics **18**(11), 79–86 (2017)
20. Wiatrak, M., Iso-Sipila, J.: Simple hierarchical multi-task neural end-to-end entity linking for biomedical text. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pp. 12–17 (2020)
21. Fu, X., Zhang, J., Meng, Z., King, I.: Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In: Proceedings of The Web Conference vol. 2020, pp. 2331–2341 (2020)

22. D’Souza, J., Ng, V.: Sieve-based entity linking for the biomedical domain. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 297–302 (2015)
23. Kang, N., Singh, B., Afzal, Z., Mulligen, E.M., Kors, J.A.: Using rule-based natural language processing to improve disease normalization in biomedical text. J. Am. Med. Inform. Assoc. **20**(5), 876–881 (2013)
24. Leal, A., Martins, B., Couto, F.M.: Ulisboa: recognition and normalization of medical concepts. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 406–411 (2015)
25. Buyko, E., Tomanek, K., Hahn, u.: 2007. resolution of coordination ellipses in biological named entities using conditional random fields. In: In PACLING 2007 - Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pp. 163–171 (2007)
26. Savova, G.K., Coden, A.R., Sominsky, I.L., Johnson, R., Ogren, P.V., Groen, P., Chute, C.G.: Word sense disambiguation across two domains: biomedical literature and clinical notes. J. Biomed. Inform. **41**(6), 1088–1100 (2008)
27. Stevenson, M., Guo, Y., Alamri, A., Gaizauskas, R.: Disambiguation of biomedical abbreviations (2009)
28. Gaudan, S., Kirsch, H., Rebholz-Schuhmann, D.: Resolving abbreviations to their senses in medline. Bioinform. **21**(18), 3658–3664 (2005)
29. Xu, J., Lee, H.-J., Ji, Z., Wang, J., Wei, Q., Xu, H.: Uth_Ccb system for adverse drug reaction extraction from drug labels at tac-Adr 2017. In: TAC (2017)
30. Leaman, R., Lu, Z.: Taggerone: joint named entity recognition and normalization with semi-markov models. Bioinformatics **32**(18), 2839–2846 (2016)
31. Luo, Y., Song, G., Li, P., Qi, Z.: Multi-task medical concept normalization using multi-view convolutional neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
32. Schumacher, E., Mulyar, A., Dredze, M.: Clinical concept linking with contextualized neural representations (2020)
33. Xu, D., Zhang, Z., Bethard, S.: A generate-and-rank framework with semantic type regularization for biomedical concept normalization, pp 8452–8464 (2020)
34. Ji, Z., Wei, Q., Xu, H.: Bert-based ranking for biomedical entity normalization (2019)
35. Zhao, S., Liu, T., Zhao, S., Wang, F.: A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 817–824 (2019)
36. Niu, J., Yang, Y., Zhang, S., Sun, Z., Zhang, W.: Multi-task character-level attentional networks for medical concept normalization. Neural. Process. Lett. **49**(3), 1239–1256 (2019)
37. Deng, P., Chen, H., Huang, M., Ruan, X., Xu, L.: An ensemble cnn method for biomedical entity normalization. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pp. 143–149 (2019)
38. Murty*, S., Verga*, P., Vilnis, L., Radovanovic, I., McCallum, A.: Hierarchical losses and new resources for fine-grained entity typing and linking. arXiv (2018)
39. Mondal, I., Purkayastha, S., Sarkar, S., Goyal, P., Pillai, J., Bhattacharyya, A., Gattu, M.: Medical entity linking using triplet network. arXiv preprint. arXiv:2012.11164 (2020)
40. Fakhraei, S., Mathew, J., Ambite, J.L.: NSEEN: neural semantic embedding for entity normalization. In: Machine Learning and Knowledge Discovery In, pp. 665–680. Springer (2019)
41. Angell, R., Monath, N., Mohan, S., Yadav, N., McCallum, A.: Clustering-based inference for zero-shot biomedical entity linking. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies
42. Vretinaris, A., Lei, C., Efthymiou, V., Qin, X., Özcan, F.: Medical entity disambiguation using graph neural networks. In: Proceedings of the 2021 International Conference on Management of Data, pp. 2310–2318 (2021)
43. Kate, R.J.: Normalizing clinical terms using learned edit distance patterns. J. Am. Med. Inform. Assoc. **23**(2), 380–386 (2015)
44. Lee, K., Hasan, S.A., Farri, O., Choudhary, A., Agrawal, A.: Medical concept normalization for online user-generated texts. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI) (2017)
45. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium, p. 17, Medical Informatics Association (2001)
46. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. **17**(5), 507–513 (2010)
47. Dogan, R.I., Lu, Z.: An inference method for disease name normalization. In: Information Retrieval and Knowledge Discovery in Biomedical Text, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012. AAAI Technical Report (2012)

48. Wermter, J., Tomanek, K., Hahn, U.: High-performance gene name normalization with GeNo. Bioinformatics **25**(6), 815–821 (2009)
49. Zhang, W., Tan, C.L., Su, J., Wang, W.T.: Entity linking leveraging automatically generated annotation. In: The 23rd International Conference on Computational Linguistics, Beijing, pp. 1290–1298. Institute for Infocomm Research (2010)
50. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12**(ARTICLE), 2493–2537 (2011)
51. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv (2013)
52. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018)
53. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinform. **36**(4), 1234–1240 (2020)
54. Wei, Q., Ji, Z., Si, Y., Du, J., Wang, J., Tiryaki, F., Wu, S., Tao, C., Roberts, K., Xu, H.: Relation extraction from clinical narratives using pre-trained language models. In: AMIA Annual Symposium Proceedings, vol. 2019, p. 1236. American Medical Informatics Association (2019)
55. Vashishth, S., Newman-Griffis, D., Joshi, R., Dutt, R., Rosé, C.P.: Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. J. Biomed. Inform. **121**, 103880 (2021)
56. Sung, M., Jeon, H., Lee, J., Kang, J.: Biomedical entity representations with synonym marginalization. arXiv (2020)
57. Miftahutdinov, Z., Tutubalina, E.: Kfu at Clef Ehealth 2017 Task 1: Icd-10 coding of english death certificates with recurrent neural networks. In: CLEF (Working Notes) (2017)
58. Cao, N.D., Izacard, G., Riedel, S., Petroni, F.: Autoregressive entity retrieval. coRR (2020)
59. Rajani, N.F., Bornea, M., Barker, K.: Stacking with auxiliary features for entity linking in the medical domain. In: BioNLP 2017, pp. 39–47 (2017)
60. Mrini, K., Nie, S., Gu, J., Wang, S., Sanjabi, M., Firooz, H. (2022)
61. Chen, Z., Ji, H.: Collaborative ranking: a case study on entity linking. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 771–781 (2011)
62. Chisholm, A., Hachey, B.: Entity disambiguation with web links. Transactions of the Association for Computational Linguistics **3**, 145–156 (2015)
63. Lazic, N., Subramanya, A., Ringgaard, M., Pereira, F.: Plato: a selective context model for entity resolution. Trans. Assoc. Comput. Linguist. **3**, 503–515 (2015)
64. Yamada, I., Shindo, H., Takeda, H., Takefuji, Y. (2016)
65. Han, X., Sun, L., Zhao, J.: Collective entity linking in web text: a graphbased method. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 765–774 (2011)
66. Cassidy, T., Ji, H., Ratinov, L.-A., Zubiaga, A., Huang, H.: Analysis and Enhancement of Wikification for Microblogs with Context Expansion. In: COLING, vol. 12, pp. 441–456 (2012)
67. He, Z., Liu, S., Song, Y., Li, M., Zhou, M., Wang, H.: Efficient collective entity linking with stacking. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 426–435 (2013)
68. Cheng, X., Roth, D.: Relational inference for wikification. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1787–1796 (2013)
69. Durrett, G., Klein, D.: A joint model for entity analysis: coreference, typing, and linking. Trans. Assoc. Comput. Linguist. **2**, 477–490 (2014)
70. Huang, H., Cao, Y., Huang, X., Ji, H., Lin, C.-Y.: Collective tweet wikification based on semi-supervised graph regularization. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 380–390 (2014)
71. Zheng, J.G., Howsmon, D., Zhang, B., Hahn, J., McGuinness, D., Hendler, J., Ji, H.: Entity linking for biomedical literature. BMC Med. Inform. Decis. Making **15**(1), 1–9 (2015)
72. Pujary, D., Thorne, C., Aziz, W.: Disease Normalization with graph embeddings. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) Intelligent Systems and Applications, pp. 209–217. Springer, Cham (2021)
73. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. **3**0 (2017)
74. Schlichtkrull, M., Kipf, T.N., Bloem, P., Berg, R.v.d., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: European Semantic Web Conference, pp. 593–607. Springer (2018)
75. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), 267–270 (2004)

76. Mohan, S., Li, D.: Medmentions: a large biomedical corpus annotated with UMLS concepts. arXiv (2019)
77. Ji, H., Nothman, J., Hachey, B., Florian, R.: Overview of Tac-Kbp2015 Tri-Lingual Entity Discovery and Linking. In: TAC (2015)
78. Afzal, Z., Akhondi, S.A., Haagen, H., Mulligen, E.M., Kors, J.A.: Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms. In: CLEF (Working Notes) (2015)
79. Van Mulligen, E.M., Afzal, Z., Akhondi, S., Vo, D., Kors, J.: Erasmus Mc at Clef Ehealth 2016: concept recognition and coding in French texts. In: CEUR Workshop Proceedings, pp. 171–178 (2016)
80. Jiang, J., Guan, Y., Zhao, C.: Wi-Enre in Clef Ehealth Evaluation Lab 2015: clinical named entity recognition based on Crf. In: CLEF (Working Notes) (2015)
81. Roller, R., Kittner, M., Weissenborn, D., Leser, U.: Cross-lingual candidate search for biomedical concept Normalization. arXiv (2018)
82. Liu, F., Vulić, I., Korhonen, A., Collier, N.: Learning Domain-specialised representations for cross-Lingual. Biomedical Entity Linking. arXiv (2021)
83. Borchert, F.: Schapranow, M.-P. Spanish biomedical entity linking with pre-trained transformers and cross-lingual candidate retrieval, Hpi-dhc@ bioasq distemist (2022)
84. Lai, T., Ji, H., Zhai, C.: Bert might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks. arXiv preprint. arXiv:2109.02237 (2021)
85. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. CoRR (2020)
86. Chen, L., Varoquaux, G., Suchanek, F.M.: A lightweight neural model for biomedical entity linking. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 12657–12665 (2021)
87. Bhowmik, R., Stratos, K., Melo, G.: Fast and effective biomedical entity linking using a dual encoder. arXiv preprint arXiv:2103.05028 (2021)
88. Ye, D., Lin, Y., Li, P., Sun, M., Liu, Z.: A simple but effective pluggable entity lookup table for pre-trained language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pp. 523–529 (2022)
89. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: Biocreative V CDR task corpus: a resource for chemical disease relation extraction. Database **2**016 (2016)
90. Doğan, R.I., Leaman, R., Lu, Z.: Ncbi disease corpus: a resource for disease name recognition and concept normalization. J. Biomed. Inform. **47**, 1–10 (2014)
91. Davis, A.P., Wiegers, T.C., Rosenstein, M.C., Mattingly, C.J.: MEDIC: a practical disease vocabulary used at the comparative toxicogenomics database. Database **2012** (2012)
92. Pradhan, S., Elhadad, N., South, B.R., Martinez, D., Christensen, L.M., Vogel, A., Suominen, H., Chapman, W.W., Savova, G.K.: Task 1: Share/Clef Ehealth Evaluation Lab 2013. In: CLEF (Working Notes), Vol. 1179 (2013)
93. Basaldella, M., Liu, F., Shareghi, E., Collier, N.: COMETA: a corpus for medical entity linking in the social media. arXiv (2020)
94. Wright, D., Katsis, Y., Mehta, R., Hsu, C.-N.: Normco: deep disease normalization for biomedical knowledge base construction. In: Automated Knowledge Base Construction (AKBC) (2019)
95. Varma, M., Orr, L., Wu, S., Leszczynski, M., Ling, X., Ré, C.: Cross-domain data integration for named entity disambiguation in biomedical text. arXiv preprint. arXiv:2110.08228 (2021)
96. Abdurxit, M., Tohti, T., Hamdulla, A.: An efficient method for biomedical entity linking based on inter- and intra-entity attention. Appl. Sci. **12**(6), 3191 (2022)
97. Dong, H., Suárez-Paniagua, V., Zhang, H., Wang, M., Casey, A., Davidson, E., Chen, J., Alex, B., Whiteley, W., Wu, H.: Ontology-based and Weakly Supervised Rare Disease Phenotyping from Clinical Notes. arXiv (2022)
98. Du, C., Popat, K., Martin, L., Petroni, F.: Entity tagging: extracting entities in text without mention supervision. coRR (2022)
99. Ayoola, T., Tyagi, S., Fisher, J., Christodoulopoulos, C., Pierleoni, A.: RefinED: an efficient zero-shot-capable approach to end-to-end entity linking. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, pp. 209–220. Association for Computational Linguistics, Hybrid: Seattle, Washington + Online (2022)
100. Dong, S., Miao, X., Liu, P., Wang, X., Cui, B., Li, J. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE), pp. 1754–1766 (2022)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Jiyun Shi[1] · Zhimeng Yuan[1] · Wenxuan Guo[1] · Chen Ma[2] · Jiehao Chen[3] · Meihui Zhang[1]

Jiyun Shi
shijiyun@bit.edu.cn

Zhimeng Yuan
yuanzhimeng2002@gmail.com

Wenxuan Guo
johnsonguo@bit.edu.cn

Chen Ma
chenma@cityu.edu.hk

Jiehao Chen
chenjiehao@china-aii.com

[1] School of Computer Science and Technology, Beijing Institute of Technology,
Beijing, 100081, China

[2] Department of Computer Science, City University of Hong Kong, Hong Kong, 999077, China

[3] Institute for Data Application and Management, China Academy of Industrial Internet,
Beijing, 100102, China