Medical Entity Disambiguation with Medical Mention Relation and Fine-grained Entity Knowledge

Wenpeng Lu[1,4], Guobiao Zhang[1,4], Xueping Peng[2], Hongjiao Guan[1,4,*], Shoujin Wang[3]
[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education,
Shandong Computer Science Center (National Supercomputer Center in Jinan)
Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia
[3]Data Science Institute, University of Technology Sydney, Sydney, Australia
[4]Shandong Provincial Key Laboratory of Computer Networks,
Shandong Fundamental Research Center for Computer Science, Jinan, China
{wenpeng.lu,hongjiao.guan}@qlu.edu.cn, zgb_stubbron@163.com
{xueping.peng,shoujin.wang}@uts.edu.au

Abstract

Medical entity disambiguation (MED) plays a crucial role in natural language processing and biomedical domains, which is the task of mapping ambiguous medical mentions to structured candidate medical entities from knowledge bases (KBs). However, existing methods for MED often fail to fully utilize the knowledge within medical KBs and overlook essential interactions between medical mentions and candidate entities, resulting in knowledge- and interaction-inefficient modeling and suboptimal disambiguation performance. To address these limitations, this paper proposes a novel approach, MED with Medical Mention Relation and Fine-grained Entity Knowledge (MMR-FEK). Specifically, MMR-FEK incorporates a mention relation fusion module and an entity knowledge fusion module, followed by an interaction module. The former employs a relation graph convolutional network to fuse mention relation information between medical mentions to enhance mention representations, while the latter leverages an attention mechanism to fuse synonym and type information of candidate entities to enhance entity representations. Afterwards, an interaction module is designed to employ a bidirectional attention mechanism to capture interactions between medical mentions and entities to generate the matching representation. Extensive experiments on two publicly available real-world datasets demonstrate MMR-FEK's superiority over state-of-the-art(SOTA) MED baselines across all metrics. Our source code is publicly available.

Keywords: Medical Entity Disambiguation, Word Sense Disambiguation, Fine-grained Entity Knowledge, Knowledge Discovery, Text Mining, Bidirectional Attention Mechanism

1. Introduction

The continuous development in the healthcare field has led to a significant increase in the volume of biomedical texts, including electronic health records (EHRs) and biomedical literature. It is imperative to harness the vast knowledge embedded in these records to deliver high-quality information that supports clinical decision-making (Ahmadi and Nopour, 2022; Xu et al., 2023). However, it is important to note that many medical concepts can have mentions that are remarkably similar, making them difficult to differentiate, even for biomedical experts. Incorrectly disambiguating these mentions can lead to a misinterpretation of the overall context, thereby posing a significant risk in healthcare-related decision-making.

Medical entity disambiguation (MED) refers to the task of mapping medical mentions in medical text documents to their corresponding entities in a knowledge base (KB) like the unified medical language system (UMLS) (Bodenreider, 2004). For example, as shown in Figure 1, let's consider the following sentence: "Insulinum is an important factor in the treatment of DM." The abbreviation "DM" is a medical mention that could refer to the entities "diabetes mellitus" or "diabetic microangiopathy" in UMLS. The MED system is required to accurately map the mention "DM" to the appropriate entity "diabetes mellitus". Furthermore, MED has diverse applications in various down-

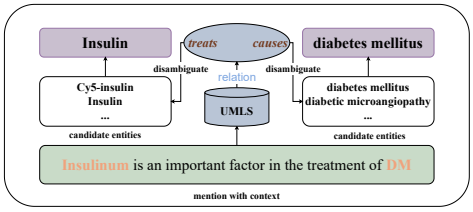

Figure 1: Example of MED with medical mention relation information. "treats" and "causes" denote the relation information between the medical mentions "Insulinum" and "DM", which are retrieved from UMLS. According to the relations, "Insulinum" and "DM" can be disambiguated to "Insulin" and "diabetes mellitus".

*Corresponding author

LREC-COLING 2024, pages 11148–11158

stream tasks, including medical-related decision making (Wang et al., 2022), predictive modeling (Dhamala et al., 2023), medical information extraction (Agrawal et al., 2022), and question answering (Eslami et al., 2023; Lu et al., 2024).

In recent years, graph neural networks (GNNs) have gained significant successes in enhancing the performance of MED (Al-Sabri et al., 2022; Vretinaris et al., 2021; Yang et al., 2023b). One noteworthy approach is ED-GNN, which treats MED as a graph matching problem, leveraging GNNs to disambiguate medical mentions (Vretinaris et al., 2021). However, this approach primarily focuses on utilizing GNNs to model the contextual information of medical mentions and candidate entities, while overlooking the vast amount of knowledge information contained in the medical KB, such as synonyms and type information of entities. This inevitably leads to the suboptimal performance.

In order to mitigate the aforementioned challenges, researchers have attempted to incorporate knowledge information from medical KB into MED models. LATTE introduces the concept of latent type modeling for entities and their mentions, enhancing MED by incorporating fine-grained latent type information about mentions and candidate entities (Zhu et al., 2020). SAPBERT adopts a pretraining scheme that self-aligns the representations of biomedical entities, exploiting fine-grained semantic information through the alignment of synonymous entities to enhance entity representations (Liu et al., 2021). Grissette and Nfaoui (2022) combine general knowledge and medical synonyms knowledge to build a semi-supervised MED model. Although existing models have improved MED performance, they often focus on specific type or synonyms information while neglecting more relation information in KBs. Moreover, none of the above methods fully consider the interaction between medical mentions and candidate entities. These limitations hinder the improvement of MED models.

To this end, we propose a novel approach to improve medical entity disambiguation with medical mention relation and fine-grained entity knowledge (MMR-FEK). Specifically, MMR-FEK first utilizes a mention relation fusion module to extract relation information from UMLS between medical mentions, and then the mention relation fusion module uses RGCN to integrate the relation information to enhance the semantic representation of medical mentions. Subsequently, MMR-FEK utilizes an entity knowledge fusion module to incorporate the synonym and type information of candidate entities to enhance the semantic representation of candidate entities. Then, MMR-FEK leverages the bidirectional attention mechanism in an interaction module to capture the interactions between medical mentions and entities, considering both entity-to-mention and mention-to-entity interactions, to generate the matching representation. Finally, MMR-FEK designs a matching module to predict the matching score to judge the right candidate entity.

Accordingly, this paper makes the following main contributions:

- We propose a novel approach for medical entity disambiguation with medical mention relation and fine-grained entity knowledge (MMR-FEK). In contrast to existing works, our approach mines more additional relation knowledge between mentions and more fine-grained entity knowledge to enhance their representations, and captures more interactions between mentions and entities. To the best of our knowledge, this is the first work that fully and simultaneously leverages the medical mention relation and fine-grained entity knowledge to improve medical entity disambiguation.

- We devise three core modules for MMR-FEK, including a mention relation fusion module and an entity knowledge fusion module, followed by an interaction module. The former first leverages RGCN to effectively fuse mention relation information between medical mentions. Simultaneously, the latter utilizes an attention mechanism to fuse fine-grained synonym and type information of candidate entities. Subsequently, an interaction module is designed to fully capture the interactions between medical mentions and entities to generate the matching representation.

- Extensive experiments on two publicly available real-world datasets demonstrate MMR-FEK's superiority over state-of-the-art MED baselines across all metrics.[1]

2. Related Work

In this section, we first introduce entity disambiguation work in the general domain, and then focus on MED work in the biomedical domain.

2.1. Entity disambiguation

Entity disambiguation is a fundamental task in natural language processing, which has attracted significant interest from researchers in recent years. Traditional approaches for entity disambiguation typically rely on rule-based approaches (Cucerzan, 2007), where the similarities between mentions and entities are computed using manually defined

[1]All source code and datasets of this paper can be obtained from https://github.com/Stubborn-z/MMR-FEK.

rules and feature engineering. However, the process of manually creating rules is time-consuming, labor-intensive, and expensive. In comparison to rule-based methods, machine learning-based approaches for entity disambiguation exhibit superior flexibility and generalization capabilities, resulting in higher accuracy. For instance, Francis-Landau et al. (2016) employ convolutional neural networks (CNNs) to capture the semantic similarities between source document mentions and potential target entities. Nevertheless, these methods face limitations in utilizing semantic knowledge information in KBs to differentiate similar and ambiguous entities. To harness semantic knowledge from KBs more effectively, recent works focus on incorporating knowledge into deep learning-based approaches, yielding impressive performance (Zhang et al., 2022a). For example, K-NED introduces both factual and conceptual knowledge graphs into the entity disambiguation task, which offers additional knowledge to enhance the semantic matching between mentions and candidate entities (Feng et al., 2020). Additionally, ExtEnD augments title information of Wikipedia with the description information from Wikidata to address the limitation of insufficient mention representation in the extractive entity disambiguation model (Procopio et al., 2023). Although numerous entity disambiguation methods exist in the general domain, they cannot be directly applied to the biomedical domain.

2.2. Medical Entity disambiguation

Due to significant disparities in linguistic features and KBs between general and biomedical domains, MED poses a significant challenge for researchers. Several recent studies have demonstrated that incorporating auxiliary information, such as type or specific semantic knowledge, alongside pretraining, can enhance the performance of the model in diverse medical NLP tasks (Lee et al., 2022; Zhang et al., 2022b). To fully utilize the knowledge in KBs, LATTE improves MED by modeling latent fine-grained type information of medical mentions and candidate entities, employing an attention-based mechanism to rank candidate entities (Zhu et al., 2020). SAPBERT adopts a self-alignment pretraining scheme to fine-tune BERT on synonyms extracted from UMLS, enhancing the semantic representation of medical entities (Liu et al., 2021). Additionally, Zhu et al. (2021) utilizes a two-stage algorithm to enhance entity representation based on prompt learning, where a coarser-grained retrieval independently embeds the surface forms of mentions and entities in the first stage while a fine-grained encoder re-ranks each candidate in the second stage . Beyond the context within the current documents, B-LBConA further involves external documents, which focuses on capturing dependencies with other medical documents and efficiently extracts hidden information in medical contextual texts (Yang et al., 2023a). Although the aforementioned methods make use of knowledge information to some extent, aiming to alleviate the issue of inadequate representation of medical mentions and candidate entities, they primarily rely on a single type of knowledge and do not effectively integrate the knowledge embedded within medical KBs, which inadequately models the semantic representation of medical mentions and the semantic representation of candidate entities. Additionally, there is a lack of explicit interaction between medical mentions and candidate entities, which limits the performance of existing methods.

3. Methodology

In this section, we first introduce the problem statement of MED. Then, we provide a detailed explanation of our proposed model, i.e., MED with medical mention relation and fine-grained entity knowledge (MMR-FEK).

3.1. Problem Statement

Given the biomedical context $\mathbb{C} = \{x_1^c, \ldots, x_l^c\}$ consisting of l tokens with N mentions $\mathbb{M} = \{m_1, \ldots, m_N\}$ and a set of M entities $\mathbb{E} = \{e_1, \ldots, e_M\}$ in a target biomedical KB. The task of MED involves associating each medical mention m in the context with a corresponding target entity e from the set $\mathbb{E}$. Each medical mention in the context may span across one or more tokens, indicated by a pair of start and end indices (i, j), denoted as $m_a = \{x_i^c, \ldots, x_j^c\}$.

3.2. Model Architecture

The overall architecture of our model is illustrated in Figure 2, which contains five main components: *embedding module*, *mention relation fusion module*, *entity knowledge fusion module*, *interaction module*, and *matching module*. First, given a context containing ambiguous medical mentions and candidate entities, the *embedding module* employs the medical pre-trained language model BioBERT to obtain their contextualized representations respectively. Then, the *mention relation fusion module* retrieves all relations among mentions from UMLS, and employs a relation graph convolutional network (RGCN) to generate relation representations, which are further fused with the contextualized representations to generate the enhanced semantic representation of medical mentions. Simultaneously, the *entity knowledge fusion module* utilizes an attention mechanism to combine fine-grained synonym and type information together,

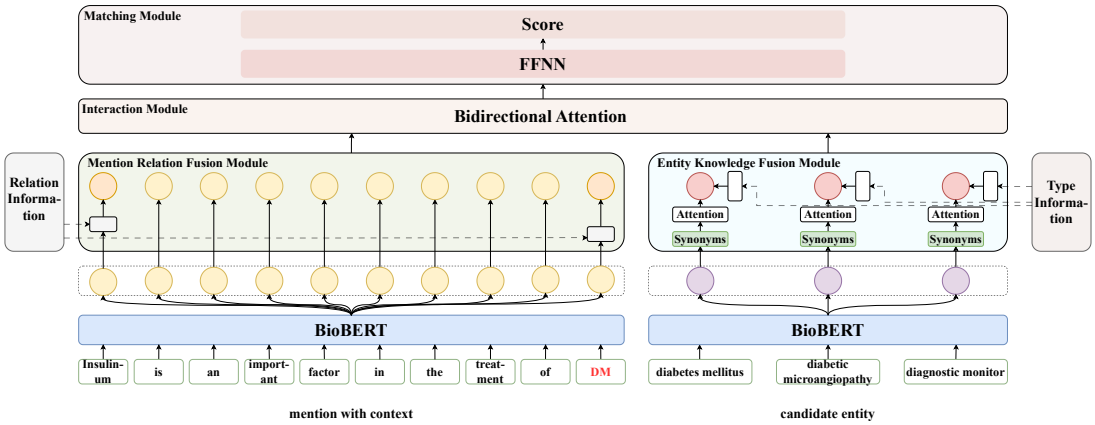

Figure 2: MMR-FEK is built on five core components: the *embedding module*, *mention relation fusion module*, *entity knowledge fusion module*, *interaction module*, and *matching module*. A detailed explanation can be found in Section 3.2.

generating the enhanced semantic representation of candidate entities. Subsequently, the *interaction module* fully captures the interactions between medical mentions and candidate entities with bidirectional attention mechanism to generate their matching representations. Finally, the *matching module* predicts the matching score of each candidate entity, and selects the one with highest score as the right entity. Next, we will introduce these five modules in detail one by one.

3.2.1. Embedding Module

As is well known, it has been demonstrated that large-scale medical pre-trained language models (PLMs)(Michalopoulos et al., 2021),(Lee et al., 2020) are highly effective in various medical tasks, which enables a more profound comprehension of medical texts. Therefore, we adopt the medical PLM BioBERT as the encoder in the embedding module (Lee et al., 2020). Specifically, for the input biomedical text $\mathbb{C} = \{x_1^c, \ldots, x_l^c\}$ with N mentions and each candidate entity $e_i = \{x_1^e, \ldots, x_u^e\}$ consisting of u tokens, we utilize BioBERT to obtain contextualized representation of medical mentions as well as representations of candidate entities, as follows:

$$\mathbf{H}^c = \{\mathbf{h}_1^c, \ldots, \mathbf{h}_l^c\} = \text{BioBERT}\{x_1^c, \ldots, x_l^c\} \quad (1)$$

$$\mathbf{H}^e = \{\mathbf{h}_1^e, \ldots, \mathbf{h}_u^e\} = \text{BioBERT}\{x_1^e, \ldots, x_u^e\} \quad (2)$$

where $\mathbf{H}^c \in \mathbb{R}^{d_h}$ and $\mathbf{H}^e \in \mathbb{R}^{d_h}$ denote the contextualized representation of medical mentions and candidate entity, and d_h refers to the hidden dimension of BioBERT. For each medical mention $m_a = \{x_i^c, \ldots, x_j^c\}$, represented by BioBERT encoding as $\mathbf{m}_a = \{\mathbf{h}_i^c, \ldots, \mathbf{h}_j^c\}$.

3.2.2. Mention Relation Fusion Module

It is obvious that utilizing more relation information between medical mentions is important for enhancing their representation, which directly affect the disambiguation ability. Thus, we devise a mention relation fusion module to fuse relation information to enhance mention representation.

As RGCN has demonstrated powerful ability to model relation structures in graphs (Chen et al., 2022), we adopt RGCN to encode relations to generate comprehensive and accurate representations of relation information between medical mentions, as shown in Figure 3. Initially, all medical mentions $\mathbb{M}$ within the context $\mathbb{C}$, along with mention relations $\mathbb{R}$, are retrieved from the UMLS knowledge graph. Subsequently, these relations $\mathbb{R}$ are encoded with RGCN, as below:

$$\mathbb{R} = \text{GE}(\mathbb{M}) \quad (3)$$

$$\mathbf{h}_i^l = \sigma\Big(\sum_{r\in\mathbb{R}}\sum_{j\in N_i^r}\frac{1}{c_i^r}\mathbf{W}_r^{(l-1)}\mathbf{h}_j^{(l-1)} + \mathbf{W}_0^{(l-1)}\mathbf{h}_i^{l-1}\Big) \quad (4)$$

$$\mathbf{H}^r = \{\mathbf{h}_1^r, \ldots, \mathbf{h}_g^r\} \quad (5)$$

where GE represents an embedded lookup table for knowledge graphs, $\mathbf{h}_i^l$ refers to the hidden state of the relation r of the i-th mention on l-th layer of RGCN. σ is the sigmoid activation function. $\mathbf{W}_r^{(l-1)}$ and $\mathbf{W}_0^{(l-1)}$ are the learnable parameter matrix. N_i^r denotes the set of neighbor indices for the i-th medical mention according to $\mathbb{R}$. c_i^r is a normalization constant. $\mathbf{h}_g^r$ denotes the hidden state of the g-th mention relation r after passing through RGCN. $\mathbf{H}^r$ denotes the relation representations.

Afterwards, as depicted in the left panel of Figure 2, we further fuse the aforementioned relation representation $\mathbf{H}^r$ between medical mentions with the

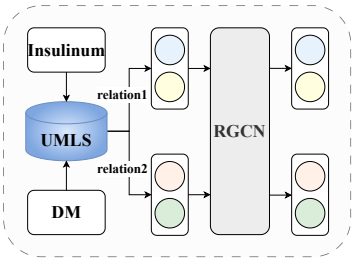

Figure 3: The schematic diagram for encoding relation information with RGCN.

contextualized representation of medical mentions $\mathbf{H}^c$ through concatenation and linear operations, generating the semantic representations of medical mentions, as below:

$$\mathbf{H}^{c,r} = \mathbf{W}\left[\mathbf{H}^c; \mathbf{H}^r\right] + \mathbf{b} \tag{6}$$

where $\mathbf{H}^{c,r}$ denotes the enhanced semantic representation of medical mentions, $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters.

3.2.3. Entity Knowledge Fusion Module

To obtain an optimal entity representation, it is essential to incorporate more entity knowledge. Both synonyms and type information can enhance the representations of entities, facilitating medical entity disambiguation. Thus, we devise an entity knowledge fusion module to incorporate synonyms and type information.

As shown in the right panel of Figure 2, we first search candidate entities and their synonyms from UMLS and encode them with BioBERT to obtain their respective representations, denoted as $\mathbf{H}^e$ and $\mathbf{H}^s$. Then, we incorporate the synonym information representation $\mathbf{H}^s$ into the candidate entities representation $\mathbf{H}^e$ through the attention mechanism (Vaswani et al., 2017), as below:

$$\begin{aligned}\mathbf{H}^{e,s} &= \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{\mathbf{d}_k}})\mathbf{V}\end{aligned} \tag{7}$$

where $\mathbf{Q}$ represents the query matrix of candidate entities, $\mathbf{K}$ represents the key matrix of candidate entity synonyms, and $\mathbf{V}$ represents the value matrix of candidate entities. $\mathbf{H}^{e,s}$ is the representation of the candidate entity fused with synonym knowledge.

Afterwards, we retrieve the type of candidate entities from UMLS and encode it with BioBERT to obtain type information representation $\mathbf{H}^t$. We further incorporate type information representation $\mathbf{H}^t$ into $\mathbf{H}^{e,s}$ to generate the enhanced semantic representation of candidate entities $\mathbf{H}^{e,s,t}$ through a concatenation operation, as below:

$$\mathbf{H}^{e,s,t} = [\mathbf{H}^{e,s}; \mathbf{H}^t] \tag{8}$$

3.2.4. Interaction Module

To accurately evaluate the matching degree among medical mentions and candidate entities, it is critical to model the semantic interactive features between them. Therefore, we devise an interactive module with a bidirectional attention mechanism to capture the sophisticated interactive features (Seo et al., 2017).

Specifically, once we obtain the enhanced semantic representation of medical mention $\mathbf{H}^{c,r}$ in Equation (6) and the enhanced semantic representation of candidate entity $\mathbf{H}^{e,s,t}$ in Equation (8), we first calculate their similarity matrix $\mathbf{S}$, and then model their interactions through the bidirectional attention mechanism. The element $\mathbf{s}_{ij}$ in the matrix is computed as below:

$$s_{ij} = \mathbf{w}_a^{\mathrm{T}} \cdot [\mathbf{h}_i^{c,r}; \mathbf{h}_i^{e,s,t}; \mathbf{h}_i^{c,r} \odot \mathbf{h}_i^{e,s,t}] \tag{9}$$

where s_{ij} refers to the similarity between the i-th token of the medical mention and the j-th token of the candidate entity. $\mathbf{w}_a$ is a trainable weight vector, and $\odot$ denotes the dot product.

Afterwards, we implement bidirectional attention mechanism, computing mention-to-entity attention $\mathbf{att}_i^m$ and entity-to-mention attention $\mathbf{att}_j^e$, respectively, as follows:

$$\begin{aligned}\overline{\mathbf{S}}^{\alpha} &= \mathrm{softmax}(row(\mathbf{S})) \\ \mathbf{att}_i^m &= \mathbf{h}_i^{c,r} \odot \overline{\mathbf{S}}^{\alpha}\end{aligned} \tag{10}$$

$$\begin{aligned}\overline{\mathbf{S}}^{\beta} &= \mathrm{softmax}(\max_{col}(\mathbf{S})) \\ \mathbf{att}_j^e &= \mathbf{h}_j^{e,s,t} \odot \overline{\mathbf{S}}^{\beta}\end{aligned} \tag{11}$$

where row denotes the row in the matrix, $\max_{col}$ denotes the maximum function calculated by column.

Finally, we calculate the attention vectors $\{\mathbf{a}_z\}_{z=1}^n$. All the vectors are concatenated to form the matching representation $\mathbf{A}$, as follows:

$$\mathbf{a}_z = [\mathbf{h}_z^{c,r}; \mathbf{h}_z^{e,s,t}; \mathbf{h}_z^{c,r} \odot \mathbf{att}_z^e; \mathbf{h}_z^{e,s,t} \odot \mathbf{att}_z^m] \tag{12}$$

$$\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \tag{13}$$

3.2.5. Matching Module

Once the matching representation of medical mention and candidate entity is obtained, we first predict their matching score using a multi-layer feed-forward neural network, and then select the entity with the highest score as the right one. We devise the matching module to carry out these detailed operations, described as follows:

$$\Phi' = \mathrm{ReLU}(\mathbf{W}_1 \cdot \mathbf{A} + \mathbf{b}_1) \tag{14}$$

$$\mathbf{S}(m, e) = \mathrm{Sigmoid}(\mathbf{W}_2 \cdot \Phi' + \mathbf{b}_2) \tag{15}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are the learnable weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are the bias.

3.3. Training

For training our model, we designate the golden entities as positive samples in the training process. Negative samples are generated from mismatch candidates generated by the SAPBERT (Liu et al., 2021), as well as randomly sampled negative instances from the full dataset. We employ the cross-entropy loss function to train our model with the objective of maximizing the scores of the golden target entities:

$$\mathcal{L} = -\sum_{i=1}^{n} \lambda(\hat{s} \log s_i + (1 - \hat{s}) \log(1 - s_i)) \quad (16)$$

where $\mathcal{L}$ represents the loss function, λ denotes the label weight that balances the positive and the negative samples. $\hat{s}$ represents the label of the gold entity, indicating the true entity for the given medical mention. s_i is the score of each candidate entity.

4. Experiments

In this section, extensive experiments on two publicly available real-world datasets are carried out to evaluate the performance of our MMR-FEK model by comparing it with those representative and state-of-the-art models. In addition, the importance of each individual component of MMR-FEK is investigated by an ablation study. We also provide a case study on relation knowledge and an influence analysis on the number of synonyms.

4.1. Datasets

We conduct experiments on two publicly avaliable real-world biomedical datasets: MedMentions[2] and BC5CDR[3]. The detailed statistics are shown in Table 1.

- **MedMentions** is the most popular and largest biomedical entity disambiguation dataset, which consists of 4,392 abstracts from PubMed, with over 350,000 mentions linked to UMLS concepts (Mohan and Li, 2018).
- **BC5CDR** consists of 1,500 articles from PubMed, with 4,409 annotated chemicals and 5,818 diseases (Li et al., 2016). It contains over 28,000 mentions linked to MeSH concepts, which are mapped to UMLS ones by us.

[2]https://github.com/chanzuckerberg/MedMentions

[3]http://www.biocreative.org/tasks/biocreative-v/track-3-cdr/

Dataset	Statistics	Train	Dev	Test
MedMentions	#Documents	2,635	878	879
	#Mentions	211,029	71,062	70,405
	#Entities	20,830	6,941	6,953
BC5CDR	#Documents	900	300	300
	#Mentions	17,135	5,710	5,714
	#Entities	5,489	1,830	1,830

Table 1: Statistics of datasets.

4.2. Evaluation Metrics

Since MED involves selecting the correct target entity from a set of candidate entities, it can be viewed as a ranking problem. To evaluate the proposed model, following the works of ED-GNN (Vretinaris et al., 2021) and B-LBConA (Yang et al., 2023a), we employ Precision@1, Recall@1, and F1 score as the evaluation metrics.

4.3. Baseline

To validate the effectiveness of MMR-FEK, we compare it with state-of-the-art approaches recently proposed in MED research.

- **NCEL** employs a graph convolutional network (GCN) to incorporate contextual features and coherence information to enhance MED(Cao et al., 2018).
- **BIOSYN** utilizes iterative candidate selection together with synonym marginalization techniques to optimize the representation of biomedical entity (Sung et al., 2020).
- **SAPBERT** implements a metric learning framework to learn self-align synonymous biomedical entities (Liu et al., 2021).
- **Dual Encoder** utilizes BERT-based dual encoders to disambiguate multiple mentions of biomedical concepts within a document simultaneously (Bhowmik et al., 2021).
- **Zhu** introduces a two-stage algorithm consisting of a coarser-grained retrieval and a finer-grained encoder to capture the contextual information of mentions and entities to disambiguate. (Zhu et al., 2021).
- **LATTE** improves MED by modeling the underlying fine-grained type information of medical mentions and entities (Zhu et al., 2020).
- **B-LBConA** is a MED model based on BioLinkBERT and the context-aware mechanism (Yang et al., 2023a).

4.4. Implementation Details

We implemented our model, along with all the other baselines, using the PyTorch framework, and selected BioBERT as our based model. In order to ensure a fair comparison, we fine-tuned the parameters of all models consistently on the development datasets. The data was divided into three sets: training, validation, and test sets, with a distribution of 60%, 20%, and 20% respectively. Our model was trained for 15 epochs using a single A100 GPU(40GB), which required around 10.5 hours on the MedMentions dataset and approximately 3 hours on the BC5CDR dataset. The complete training process of the model required approximately 21GB of VRAM. Compared to training, the inference stage was very fast and took approximately a few minutes. We optimized our model using AdamW (Loshchilov and Hutter, 2019). Table 2 lists some important hyperparameters of our model. Noted that all parameters were chosen based on the best performance on the development set.

Hyperparameters	Value
Optimizer	*AdamW*
Learning Rate	1e-5
Decay Rate	0.01
Lable Weight	0.4
Batch Size	64
Maximum Sequence Length	128
Number of Synonyms	6

Table 2: Hyperparameter settings of MMR-FEK.

4.5. Experimental Results

Table 3 shows the overall results of MMR-FEK and several recent SOTA models on the two datasets. According to the table 3, we have several noteworthy observations.

First, our model demonstrates superior performance compared to all other models on both datasets. Specifically, on the MedMentions dataset, MMR-FEK exhibits a significant improvement over the best-performing baseline model, i.e., B-LBConA. There is an increase of 0.5, 0.7, and 0.8 points respectively in terms of precision, recall, and F1 score. On the BC5CDR dataset, MMR-FEK significantly outperforms B-LBConA by a considerable margin. The precision, recall, and F1 scores have increased by 0.8, 0.9, and 1.2 points, respectively.

Second, among all the baselines, LATTE is the work most closely related to ours, while solely relies on fine-grained latent type information of mentions and entities. Compared to LATTE, our model exhibits significant performance improvements. This is because, while utilizing type information, we not only introduce the relation information of medical mentions but also incorporate synonym information of medical entities.

Third, B-LBConA is the best performing baseline, but it mainly focuses on capturing dependencies with other documents and interaction information between contextual texts. In contrast, our model mainly focuses on leveraging the knowledge information in the medical KBs and enhancing the interaction between medical mentions and candidate entities. The result is that our model outperforms B-LBConA, which further proves the effectiveness of our proposed method.

Model	MedMentions			BC5CDR		
	Precision	Recall	F1	Precision	Recall	F1
NCEL	OOM	OOM	OOM	65.7	67.3	65.2
BIOSYN	OOM	OOM	OOM	69.1	68.8	70.1
SAPBERT	53.1	56.4	52.3	70.2	72.9	72.4
Dual Encoder	62.9	67.4	65.6	84.8	82.1	83.0
Zhu	66.5	68.1	65.4	83.2	84.0	81.1
LATTE	88.2	86.5	85.6	88.2	87.0	86.3
B-LBConA	88.5	87.1	86.5	89.1	88.2	87.3
MMR-FEK (our)	**89.0***	**87.8***	**87.3***	**89.9***	**89.1***	**88.5***
Improvement(%)	0.5	0.7	0.8	0.8	0.9	1.2

Table 3: Comparison with baselines on two datasets. We bold the best score for each column, and underline the suboptimal one. "OOM" means out-of-memory. The improvement is calculated concerning the best-performing baseline, "*" means that the improvement is significant at $p < 0.05$.

4.6. Ablation Study

To investigate the effectiveness of a specific module or knowledge information in our model, we conduct ablation experiments by removing them one by one. The results of the ablation experiments on the two datasets are shown in Figure 4.

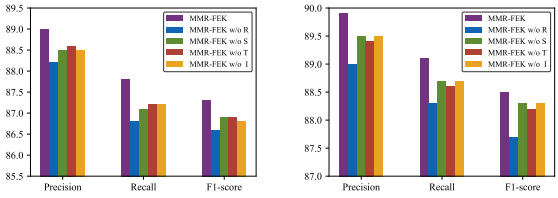

Figure 4: Ablation experimental results on MedMentions (left) and BC5CDR (right) datasets.

- **MMR-FEK w/o R** removes relation information, in which the medical mentions are encoded using BioBERT without any additional knowledge.
- **MMR-FEK w/o S** removes the synonym information of the candidate entity, in which candidate entity representations are solely concatenated with type information representations.

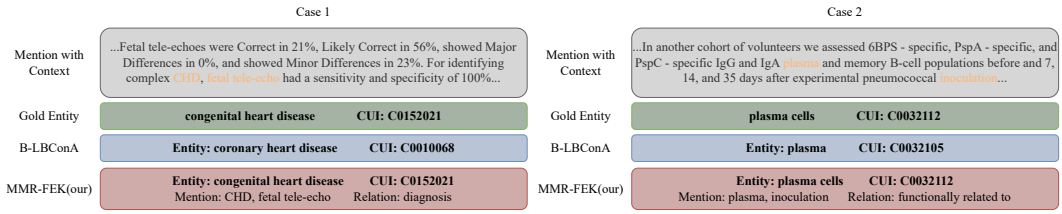

Figure 5: Case study on relation knowledge by comparing B-LBConA and MMR-FEK.

- **MMR-FEK w/o T** removes the type information, in which candidate entity representations are fused solely with synonym information representations.
- **MMR-FEK w/o I** removes the interaction module, in which medical mentions and candidate entities do not interact and are directly fed into the matching module.

According to Figure 4, we can draw several noteworthy observations. One notable observation is that removing relation information results in a significant degradation of the model's evaluation metrics on both datasets. This suggests that the relation information between medical mentions can significantly enhances the model's ability to represent medical mentions, leading to a substantial improvement in its disambiguation capabilities. Another noteworthy observation is that when the synonym information and type information of candidate entities are removed, the evaluation metrics of the model are also reduced to varying degrees. This can be attributed to the fact that the inclusion of synonym information and type information for candidate entities assists the model in improving its entity representation. A third important observation is that removing the interaction module leads to a decrease in the model's performance. This highlights the importance of reinforcing the interaction between medical mentions and candidate entities. These observations further validate the effectiveness of our approach.

4.7. Case Study on Relation Knowledge

To further demonstrate the effectiveness of relation knowledge between medical mentions, we conduct the case study. Figure 5 shows two cases from the test set of the MED dataset. In Case 1, "CHD" is the target medical mention. Due to the absence of specific "diagnosis" relation information in the two medical mentions (i.e., "CHD" and "fetal tele-echo"), the B-LBConA model incorrectly associates it with the entity "coronary heart disease". In contrast, MMR-FEK can correctly associate it with the golden entity "congenital heart disease" with the help of "diagnosis" relation information. The similar pattern is observed in Case 2, where "plasma cells" is the target medical mention. Owing to the absence of specific "functionally related to" relation information in the two medical mentions (i.e., "plasma" and "inoculation"), the B-LBConA model associates it with the wrong entity "plasma". In contrast, MMR-FEK associates it with the golden entity "plasma cells" with the help of "functionally related to" relation information.

4.8. Influence of Number of Synonyms

To investigate the influence of the number of synonyms on the performance of our approach, we conduct a quantitative analysis. The experimental results are illustrated in Figure 6. Initially, as the number of synonyms increases, the model's performance consistently improves. MMR-FEK reaches its peak performance when the number of synonyms reaches 6. This observation highlights the effectiveness of incorporating synonyms. However, it's important to note that once the number of synonyms exceeds 6, the model's performance either plateaus or even experiences a slight decrease. This could be attributed to both the limited availability of synonyms for candidate entities and the potential introduction of additional noise by certain synonyms into the model.

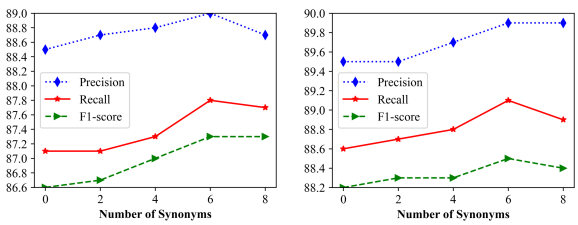

Figure 6: Influence of number of synonyms on MedMentions (left) and BC5CDR (right) datasets.

5. Conclusion and Future Work

Although recent works have achieved great successes on medical entity disambiguation (MED) task, their knowledge- and interaction-inefficient modeling, resulting in suboptimal performance. In this paper, we propose a novel MED approach with medical mention relation and fine-grained entity knowledge (MMR-FEK). In our MMR-FEK approach, we devise a mention relation fusion module to incorporate mention relation information between medical mentions to enhance mention representation. Additionally, we develop an entity knowledge fusion module to incorporate fine-grained synonym and type information of candidate entities to enhance entity representations. We also implement an interaction module to capture he interactions between mentions and entities to generate the matching representation, followed by a matching module to predict the correct candidate entity. Extensive experimental results demonstrate the effectiveness of our proposed approach on two publicly available datasets.

As future work, we plan to evaluate the performance of our model on other medical datasets to asses its generalization ability. Furthermore, generative entity disambiguation (Barba et al., 2022; Yuan et al., 2022) has attracted the interest of numerous researchers in recent years, undergoing rapid development and achieving improved performance. We plan to focus our work on generative MED while incorporating knowledge information.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.62376130), Shandong Provincial Natural Science Foundation (Grant No.ZR2022MF243, Grant NO.ZR2021QF059), Program of New Twenty Policies for Universities of Jinan (Grant No.202333008), Program of Innovation Improvement of Shandong (Grant No.2023TSGC0182).

7. Ethical Discussion

Medical entity disambiguation (MED) plays a crucial role in natural language processing and biomedical domains. It has a variety of applications in downstream medical tasks, contributing to assisting medical-related decisions, advancing medical information extraction, and facilitating accurate medical question answering. We believe that the potential for misuse of this MED technology is low. Our technology is developed using publicly available datasets, following the data use guidelines and ensuring no copyright infringement.

8. Bibliographical References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Maryam Ahmadi and Raoof Nopour. 2022. Clinical decision support system for quality of life among the elderly: An approach using artificial neural network. *BMC Medical Informatics and Decision Making*, 22(1):1–15.

Raeed Al-Sabri, Jianliang Gao, Jiamin Chen, Babatounde Moctard Oloulade, and Tengfei Lyu. 2022. Multi-view graph neural architecture search for biomedical entity and relation extraction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1221–1233.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. Extend: Extractive entity disambiguation. In *Proceedings of the 2022 Annual Meeting of the Association for Computational Linguistics*, pages 2478–2488.

Rajarshi Bhowmik, Karl Stratos, and Gerard de Melo. 2021. Fast and effective biomedical entity linking using a dual encoder. In *Proceedings of the 2021 International Workshop on Health Text Mining and Information Analysis*, pages 28–37.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, pages D267–D270.

Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018. Neural collective entity linking. In *Proceedings of the 2018 International Conference on Computational Linguistics*, pages 675–686.

Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 2022 Annual Meeting of the Association for Computational Linguistics*, pages 2974–2985.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716.

Elvisha Dhamala, BT Thomas Yeo, and Avram J Holmes. 2023. One size does not fit all: Methodological considerations for brain-based predictive modeling in psychiatry. *Biological Psychiatry*, 93(8):717–728.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics*, pages 1151–1163.

Zhifan Feng, Qi Wang, Wenbin Jiang, Yajuan Lyu, and Yong Zhu. 2020. Knowledge-enhanced named entity disambiguation for short text. In *Proceedings of the 2020 Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 2020 International Joint Conference on Natural Language Processing*, pages 735–744.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1256–1261.

H Grissette and EH Nfaoui. 2022. Semisupervised neural biomedical sense disambiguation approach for aspect-based sentiment analysis on social networks. *Journal of Biomedical Informatics*, pages 104229–104229.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pretrained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022. Overview of the rocling 2022 shared task for chinese healthcare named entity recognition. In *Proceedings of the 2022 Conference on Computational Linguistics and Speech Processing*, pages 363–368.

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

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 2019 International Conference on Learning Representations*.

Wenpeng Lu, Sibo Wei, Xueping Peng, Yi-Fei Wang, Usman Naseem, and Shoujin Wang. 2024. Medical question summarization with entity-driven contrastive learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1744–1753.

Sunil Mohan and Donghui Li. 2018. MedMentions: A large biomedical corpus annotated with umls concepts. In *Automated Knowledge Base Construction*.

Luigi Procopio, Simone Conia, Edoardo Barba, and Roberto Navigli. 2023. Entity disambiguation with entity definitions. In *Proceedings of the 2023 Conference of the European Chapter of the Association for Computational Linguistics*, pages 1297–1303.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the 2017 International Conference on Learning Representations*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Alina Vretinaris, Chuan Lei, Vasilis Efthymiou, Xiao Qin, and Fatma Özcan. 2021. Medical entity disambiguation using graph neural networks. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2310–2318.

Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. 2022. A novel framework based

on medical concept driven attention for explainable medical code prediction via external knowledge. In *Findings of the Association for Computational Linguistics*, pages 1407–1416.

Zhenran Xu, Zifei Shan, Yuxin Li, Baotian Hu, and Bing Qin. 2023. Hansel: A chinese few-shot and zero-shot entity linking benchmark. In *Proceedings of the 2023 ACM International Conference on Web Search and Data Mining*, pages 832–840.

Siyu Yang, Peiliang Zhang, Chao Che, and Zhaoqian Zhong. 2023a. B-LBConA: A medical entity disambiguation model based on bio-linkbert and context-aware mechanism. *BMC Bioinformatics*, 24(1):1–18.

Zhe Yang, Yi Huang, and Junlan Feng. 2023b. Learning to leverage high-order medical knowledge graph for joint entity and relation extraction. In *Findings of the Association for Computational Linguistics*, pages 9023–9035.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4038–4048.

Guobiao Zhang, Wenpeng Lu, Xueping Peng, Shoujin Wang, Baoshuo Kan, and Rui Yu. 2022a. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension. In *Proceedings of the 2022 International Conference on Computational Linguistics*, pages 4061–4070.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022b. Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9970–9985.

Ming Zhu, Busra Celikkaya, Parminder Bhatia, and Chandan K Reddy. 2020. LATTE: Latent type modeling for biomedical entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9757–9764.

Tiantian Zhu, Yang Qin, Qingcai Chen, Baotian Hu, and Yang Xiang. 2021. Enhancing entity representations with prompt learning for biomedical entity linking. In *Proceedings of the 2021 International Joint Conference on Artificial Intelligence*, pages 4036–4042.