Biomedical Entity Linking as Multiple Choice Question Answering

Zhenxi Lin[1], **Ziheng Zhang**[1], **Xian Wu, Yefeng Zheng**
Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China

Abstract

Although biomedical entity linking (BioEL) has made significant progress with pre-trained language models, challenges still exist for fine-grained and long-tailed entities. To address these challenges, we present BioELQA, a novel model that treats Biomedical Entity Linking as Multiple Choice Question Answering. BioELQA first obtains candidate entities with a fast retriever, jointly presents the mention and candidate entities to a generator, and then outputs the predicted symbol associated with its chosen entity. This formulation enables explicit comparison of different candidate entities, thus capturing fine-grained interactions between mentions and entities, as well as among entities themselves. To improve generalization for long-tailed entities, we retrieve similar labeled training instances as clues and concatenate the input with retrieved instances for the generator. Extensive experimental results show that BioELQA outperforms state-of-the-art baselines on several datasets.

Keywords: Biomedical entity linking, fine-grained interactions, long-tailed entities

1. Introduction

Biomedical entity linking (BioEL) refers to mapping biomedical mentions to standard entities in an ontology, such as the Unified Medical Language System (UMLS) (Bodenreider, 2004), which is essential for various downstream tasks, including automatic diagnosis (Yuan and Yu, 2021; Shi et al., 2022), drug-drug interaction prediction (Li et al., 2023; Zhang et al., 2023), and knowledge graph alignment (Xiang et al., 2021; Lin et al., 2022). Unlike entity linking in the general domain, biomedical entities often have various names, including synonyms and morphological variations, such as "motrin" is also referred to as "ibuprofen". Additionally, similar surface forms of biomedical entities can have distinct meanings, such as "Type 1 Diabetes" and "Type 2 Diabetes".

Existing methods can be mainly categorized into two types. One is the discriminative methods (Sung et al., 2020; Lai et al., 2021; Liu et al., 2021), which employed BERT-based models to encode mentions and entities into the same embedding space and disambiguated mentions by nearest neighbors, or further applied a cross-encoder to rerank top candidates by capturing fine-grained mention-entity interactions (Angell et al., 2021; Zhu et al., 2021; Xu et al., 2023). Another one is the generative methods (Yan et al., 2020; Yuan et al., 2022a,b) that directly generated linked entities using text-to-text pre-trained language models, such as BART (Lewis et al., 2020), thereby circumventing the need for negative sample mining.

However, BioEL remains challenging due to the fine-grained and long-tailed entities. First, previous methods focused on mention-entity interactions, but generally neglected fine-grained interactions between candidate entities (*i.e.*, entity-entity interactions) and struggled with ambiguous mentions of multiple candidates with similar surface forms (Xu et al., 2023). For example, when the mention “haemoglobin” is compared to two closely related candidate entities, “haemoglobin c” and “hemoglobin”, the high lexical similarity between them confuses the models, leading to potential mismatch as “haemoglobin c”. Incorporating such entity-entity interactions can provide a more holistic and nuanced representation of interactions among entities. Furthermore, the number of candidate entities can be large, and they often follow long-tailed distribution (Kim and Ganapathi, 2021), further exacerbating the difficulty of BioEL. Previous studies have demonstrated performance improvement for long-tailed entities using auxiliary information, such as entity descriptions and synonyms (Varma et al., 2021; Yuan et al., 2022b). However, collecting such data is labor-intensive, which limits its applicability.

To tackle the aforementioned challenges, we present a novel model called BioELQA, which aims at capturing the fine-grained interactions between entities and enhancing the generality of long-tailed entities. We achieve this by reframing BioEL as a Multiple Choice Question Answering (MCQA) task. Given a mention, we first use a bi-encoder retriever to efficiently retrieve top-N candidate entities from the ontology as the “answer options”. Then the mention and its answer options are all fed to a generator as a multiple choice prompt, with each answer associated with a symbol (*e.g.*, "A", "B", "C"). The prompt is structured so that the generator outputs the symbol associated with its chosen answer option. This framework enables explicit comparison and contrast among different answer options, effectively modeling the mention-entity in-

[1]Correspondence to chalerislin@tencent.com and zihengzhang@tencent.com.

LREC-COLING 2024, pages 2390–2396

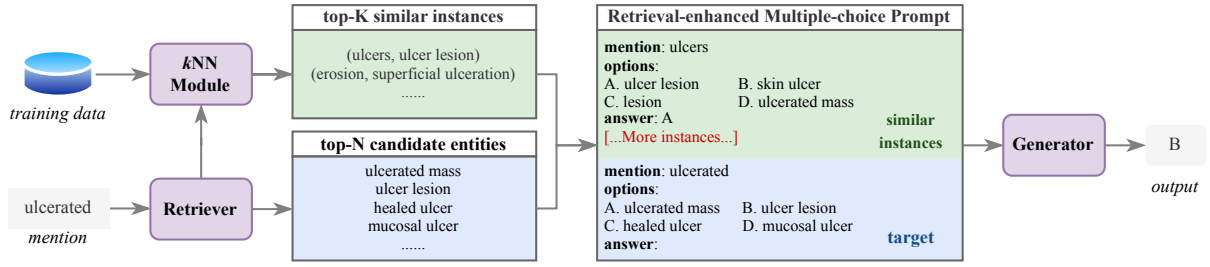

Figure 1: The overview of the proposed BioELQA.

teractions and entity-entity interactions. Besides, by directly generating answer symbols instead of entity names, we can avoid generating entities that do not exist in the ontology and reduce reliance on normalization strategies (Robinson et al., 2022). Inspired by the recent success of retrieval-augmented model (Khandelwal et al., 2020), we further leverage a kNN module to retrieve semantically similar instances from the training data to formulate a retrieval-enhanced prompt, which empowers the model to reference similar contexts as cues for prediction, to improve the robustness and generalization for long-tailed entities. We experimentally validate the effectiveness and superiority of BioELQA as it achieves state-of-the-art performance on several benchmark datasets. The source code and curated prompts will be made publicly accessible after the review.

2. Method

The goal of BioEL is to link a given mention m to its correct entity e in an ontology $\mathcal{E}$. The overall architecture of the proposed BioELQA is shown in Figure 1. Firstly, a retriever (§2.1) selects the top-N candidate entities for mention m, and a kNN module (§2.3) provides top-K similar instances from the training corpus as contextual clues. These components form a retrieval-enhanced multiple-choice prompt, which is then processed by a generator (§2.2) to produce the final answer.

2.1. Retriever

Following the previous work (Zhu et al., 2021; Xu et al., 2023), we employ a bi-encoder based on SapBERT (Liu et al., 2021) to generate dense vectors for both mentions and entities. The mention embedding $f(m)$ of mention m is formulated as:

$$f(m) = \texttt{SapBERT}(m)\texttt{[CLS]}, \quad (1)$$

where `[CLS]` denotes the special token that is used to derive a fixed-size vector. The entity embedding $f(e)$ of entity e is computed similarly. The score of a mention-entity pair (m, e) is denoted as follows:

$$s(m, e) = g(f(m), f(e)), \quad (2)$$

where g is the cosine similarity. At inference time, we precompute $f(e)$ for each $e \in \mathcal{E}$ and use Faiss (Johnson et al., 2019) for fast retrieval.

Training. We use contrastive learning to train the retriever, which aims at optimizing the agreement between true mention-entity pairs and the disagreement between false ones. The loss for each true pair (m, e) is computed as:

$$\mathcal{L}(m, e) = -\log \left(\frac{\delta(m, e)}{\delta(m, e) + \sum\limits_{e' \in \mathcal{H}(e)} \delta(m, e')} \right), \quad (3)$$

where $\delta(m, e) = \exp(s(m, e)/\tau)$, τ is a temperature hyper-parameter, and $\mathcal{H}(e) \subset \mathcal{E} \backslash \{e\}$ is a set of negatives that excludes e. We obtain $\mathcal{H}(e)$ by combining in-batch negative sampling and hard negative sampling (*i.e.*, highest-scoring incorrect entities), which has been shown beneficial for entity retrieval (Wu et al., 2020; Gao et al., 2021).

2.2. Generator

Multiple-Choice Prompting (MCP). Given m and top-N candidate entities from the retriever, we reformulate BioEL as a multiple-choice question answering task. Specifically, it involves selecting the correct answer from a set of answer options (*i.e.*, candidate entities) based on the question (*i.e.*, mention). There are two ways for answer generation: 1) the use of closed prompts (Xu et al., 2023) that fill in the blanks of a sentence through masked language modeling; 2) the use of generative models (Yuan et al., 2022a,b) that directly generate entity names. However, the former relies on well-designed prompts and continuous pre-training, and the latter could generate non-existent entities and require complex normalization strategies. An appropriate multiple-choice prompt is to jointly present the question and answer options to the model and guide it to output the symbol (*e.g.*, "A") associated with its chosen answer option.

Formally, given m as the question and its corresponding answer options $O = \{o_1, o_2, ..., o_N\}$ sorted in descending order of similarity to m, we first map (m, O) to a text sequence using a simple template $\mathcal{T}(\cdot)$: "mention: m options: A. o_1 B. o_2 C. o_3 D. o_4 answer:" (when $N = 4$), as shown in Figure 1. Then the sequence is fed into a generator $\mathcal{M}$, such as T5 (Raffel et al., 2020), to produce answer symbol a, where a is the symbol associated with the correct answer. The training objective of the generator is to maximize the likelihood: $p(a|m, O) = p_{\mathcal{M}}(\mathcal{T}(m, O))$. We train the generator using a standard seq2seq objective, *i.e.*, maximizing the output sequence likelihood with teacher forcing (Sutskever et al., 2014).

In MCP, a mention and its symbol-enumerated candidate answers are all passed to the generator as a single prompt, explicitly modeling both mention-entity and entity-entity interactions. Moreover, MCP enables the generator to only predict a single token (*e.g.*, "A") instead of entity names, avoiding the generation of invalid entities. The token-associated answer predicted by the model is the final answer. The probabilities of these symbols therefore serve as a proxy for the probability of each answer.

Previous studies (Robinson et al., 2022; Pezeshkpour and Hruschka, 2023) have found that current language models are sensitive to the order of answer options that a slight change in the order can alter the model's answer. To mitigate this problem and improve robustness, we employ a simple yet effective data augmentation strategy during training. By randomly swapping the order of O, we can construct different $(\mathcal{T}(m, O), a)$ training pairs. Theoretically, each (m, O) can generate $N!$ distinct examples. Preliminary experiments indicate that augmenting each input with one additional swap is sufficient. This strategy allows the model to learn the association between symbols and answers, rather than solely memorizing answer positions.

2.3. kNN Module

Inspired by Lin et al. (2024), we introduce a kNN module to enhance the model's generalization capabilities for long-tailed entities, enabling it to reference similar instances from the entire training corpus as prediction clues.

Given m_i in the training data, there are N candidate options O_i and a true answer symbol a_i constructed by MCP (§2.2). We construct a datastore $\mathcal{D}$ by indexing the training data as a list of key-value pairs $\{(f(m_i), (m_i, O_i, a_i))\}$, where the key is the mention embedding computed in Eq. (1) and the value is a training instance for generator. Given a mention x as input, we compute its mention embedding $f(x)$ via Eq. (1), query the datastore with $f(x)$ to all keys based on cosine similarity, and obtain the top-K most similar instances $\mathcal{N} = \{(m_{i_1}, O_{i_1}, a_{i_1}), ..., (m_{i_K}, O_{i_K}, a_{i_K})\}$. Note that during training, as x is already indexed, we filter it from the retrieved results to avoid data leakage. We convert these instances $\mathcal{N}$ into text sequences using $\mathcal{T}(\cdot)$ individually, and then concatenate them together with the input sequence $\mathcal{T}(x, O)$, where O corresponds to the answer options for x. This forms a retrieval-enhanced multiple-choice prompt that is fed into the generator to generate the answer, as shown in Figure 1. Now the training objective of the generator becomes as follows:

$$p(a|x, O, \mathcal{N}) = p_{\mathcal{M}}(\mathcal{T}(m_{i_1}, O_{i_1}) \oplus a_{i_1} \oplus \ldots \oplus \mathcal{T}(m_{i_K}, O_{i_K}) \oplus a_{i_K} \oplus \mathcal{T}(x, O)), \quad (4)$$

where $\oplus$ is concatenation operation and a is the answer symbol of (x, O). This can choose informative instances for each input dynamically and provides the generator with direct evidence about the input and references to make predictions. Our approach is similar to few-shot learning (Brown et al., 2020), yet with a key distinction of our focus on supervised learning, where model parameters are fine-tuned from given instances for performance improvement.

3. Experiments

3.1. Experimental Setup

Datasets and Evaluation. Three BioEL datasets are adopted, including NCBI (Doğan et al., 2014), BC5CDR (Li et al., 2016) and COMETA (Basaldella et al., 2020), which focus on different entity types, such as diseases and chemicals. The detailed dataset statistics are listed in Table 1. We report accuracy for all methods, and the best results are in **bold** with the second best results underlined.

	NCBI	BC5CDR	COMETA
Target entities $\lvert\mathcal{E}\rvert$	14,967	268,162	350,830
Train instances	5,784	9,285	13,489
Dev instances	787	9,515	2,176
Test instances	960	9,654	4,350

Table 1: Dataset Statistics.

Implementation Details. We use `t5-base` (Raffel et al., 2020) as the generator backbone. The number of training epochs is 20 and the batch size is 16. The hyper-parameter τ is 0.01. The number of options N and the number of similar instances K are 5 and 3, respectively. We search learning rate among $\{4\times10^{-5}, 8\times10^{-5}, 1\times10^{4}, 2\times10^{-4}, 4\times10^{-4}\}$ for different datasets. We use the AdamW optimizer to update model parameters.

Baselines. We compare the proposed BioELQA against previous state-of-the-art BioEL methods,

which are classified into three categories: 1) `discriminative-based` methods that utilize bi-encoders or extend their capabilities with cross-encoders to retrieve relevant entities, including BioSyn (Sung et al., 2020), ResCNN (Lai et al., 2021), SapBERT (Liu et al., 2021), Clustering-based (Angell et al., 2021) and Prompt-BioEL (Xu et al., 2023); 2) `generative-based` methods that directly generate the linked entities, including GenBioEL (Yuan et al., 2022b) and BioBART (Yuan et al., 2022a); 3) `LLM-based` methods that have demonstrated impressive capabilities across various tasks, especially in the biomedical domain, with simple instructions (Jahan et al., 2023), including GPT3.5,[1] PaLM-2 (Anil et al., 2023), Claude-2,[2] and LLaMA-2-13b (Touvron et al., 2023). Note that the results of `LLM-based` methods are taken from Jahan et al. (2023).

3.2. Overall Results

Table 2 shows that BioELQA outperforms previous baselines on all datasets, demonstrating the effectiveness of our proposed method. Note that the strongest baseline Prompt-BioEL requires additional synonym corpora for continued pre-training, while BioELQA only leverages the existing training data without the need for additional corpora. We observe that Claude-2 outperforms other `LLM-based` methods in all datasets, but it still significantly lags behind fully supervised methods. This suggests that fine-tuning small-scale models remains an effective choice for BioEL, as they can acquire domain-specific knowledge through parameter tuning.

Models	NCBI	BC5CDR	COMETA
BioSyn	91.1	–	71.3
ResCNN	92.4	–	80.1
SapBERT	92.3	–	75.1
Clustering-based	–	91.3	–
Prompt-BioEL	<u>92.6</u>	<u>93.7</u>	<u>83.7</u>
GenBioEL	91.9	93.3	81.4
BioBART-base	89.3	93.0	79.6
BioBART-large	89.9	93.3	81.8
GPT3.5	52.2	54.9	43.5
PaLM-2	38.4	52.1	48.8
Claude-2	70.2	78.0	53.3
LLaMA-2-13b	59.2	66.5	40.7
BioELQA	**93.5**	**94.5**	**85.2**

Table 2: Comparison of different BioEL methods on three public datasets.

[1] https://platform.openai.com/docs/models/gpt-3-5
[2] https://www.anthropic.com/index/claude-2

3.3. Ablation Study

To investigate different components in BioELQA, we compare BioELQA variants without data augmentation (w/o data aug.) in §2.2 and without kNN module (w/o kNN module). We also explore replacing similar instances with randomly selected examples from the training data (random instances) and requiring the generator to predict complete entity names instead of answer symbols (generate names). We observe that removing any component resulted in varying degrees of performance degradation, thus highlighting the contributions of these components. Particularly, generating meaningful entity names is shown more challenging than generating answer symbols alone, as it requires the model to learn from the context, leading to potential performance drops.

Model	NCBI	BC5CDR	COMETA
BioELQA	**93.5**	**94.5**	**85.2**
w/o data aug.	93.1	94.4	83.4
w/o kNN module	91.3	94.1	84.3
random instances	91.5	94.0	84.6
generate names	66.9	89.1	78.5

Table 3: Ablation study on three datasets.

3.4. Case Study

Table 4 shows two long-tailed entities *uneasy* and *misty vision* to illustrate the amendment by kNN module. Without the kNN module, the model tends to link mentions to morphologically similar but incorrect entities. Nevertheless, incorporating the kNN module allows the model to leverage related instances from the training set, enabling more informative decisions by referencing these instances.

Mention	Top-2 Similar Instances	Prediction
feel uncomfortable	(felt uncomfortable, uneasy) (uncomfortable, uneasy)	$(-)$ discomfort × $(+)$ uneasy ✓
blurred vision	(blurry vision, misty vision) (blurry, blurring of visual image)	$(-)$ double vision × $(+)$ misty vision ✓

Table 4: Two test cases from COMETA with predictions made without $(-)$ or with $(+)$ kNN Module.

Apart from the long-tailed case study, we provide case analysis between BioELQA and Prompt-BioEL. Table 5 lists four cases of BioELQA predicting correctly and Prompt-BioEL predicting incorrectly. BioELQA performs better in handling long-tailed entities or entities with morphological similarities because it can retrieve answers from similar mentions in the training set as references. In contrast, Prompt-BioEL performs better in handling ambiguous mentions because it incorporates

additional context to disambiguate, while BioELQA does not consider contextual information, as shown in Table 6 where Prompt-BioEL predicts correctly and BioELQA predicts incorrectly.

Mention	Retrieved Instances	Prediction
seizure medication	(seizure meds, seizure management), (migraine treatments, migraine prophylaxis)	(Q) seizure management (P) seizure finding
sketchy	(sketchy, incomplete), (blurry, blurring of visual image)	(Q) incomplete (P) brief
rhodiola	(rhodiola, family crassulaceae), (broccoli, brassica oleracea)	(Q) family crassulaceae (P) ralstonia
lose weight	(losing weight, weight loss)	(Q) weight loss (P) weight decreasing

Table 5: Four test cases on COMETA dataset where BioELQA (Q) made correct predictions and Prompt-BioEL (P) did not.

Mention	Context	Prediction
suppositories	for some personal reasons i can't use the [E1] suppositories [/E1]	(Q) suppository (P) suppository physical object
salt	i guess maybe some [E1] salt [/E1] residue gets on the insertion site	(Q) salt water (P) sodium chloride
sublingual	maxalt comes in a [E1] sublingual [/E1] under the tongue dissolve version	(Q) sublingual (P) sublingual intended site
gum	can't really see it but there's a dark line underneath my front teeth what seems like around my [E1] gum [/E1] shown above	(Q) gum (P) gingival structure

Table 6: Four test cases on COMETA dataset where Prompt-BioEL (P) made correct predictions and BioELQA (Q) did not.

3.5. Impacts of hyper-parameters

In Figure 2(a), we observe that the accuracy initially improves and then declines as N increases. A higher N indicates a higher recall rate but also introduces more wrong options, leading to decreased performance. Similarly, in Figure 2(b), with the increase of K, accuracy increases first and decreases afterward. This suggests that allowing too many retrieved instances can potentially confuse the model and negatively impact predictions.

4. Conclusion

We propose BioELQA, a new MCQA-based method that introduces a multiple-choice prompt that allows for more comprehensive understanding of the interactions between different entities and a given mention. In addition, a kNN module that recaptures related training data as a side-by-side reminder can explicitly provide essential information to enhance performance. We experimentally validate the state-of-the-art performance of our method in several public datasets. For future work, we plan to explore introducing contextual information to BioELQA to address entity ambiguation issues.

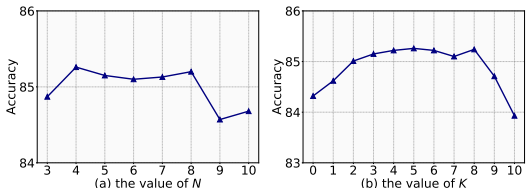

Figure 2: Impact of different hyper-parameters on the COMETA dataset.

5. Bibliographical References

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

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 technical report. *arXiv preprint arXiv:2305.10403*.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A Corpus for Medical Entity Linking in the Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *Journal of Biomedical Informatics*, 47.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. A Comprehensive Evaluation of Large Language Models on Benchmark Biomedical Text Processing Tasks. *arXiv preprint arXiv:2310.04270*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3).

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest Neighbor Machine Translation. In *International Conference on Learning Representations*.

Byung-Hak Kim and Varun Ganapathi. 2021. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In *Machine Learning for Healthcare Conference*, pages 196–208. PMLR.

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. BERT might be Overkill: A Tiny but Effective Biomedical Entity Linker based on Residual Convolutional Neural Networks. In *Findings of the Association for Computational Linguistics: EMNLP*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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*.

Xinhang Li, Zhaopeng Qiu, Xiangyu Zhao, Yong Zhang, Chunxiao Xing, and Xian Wu. 2023. REST: Drug-Drug Interaction Prediction via Reinforced Student-Teacher Curriculum Learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 1278–1287. Association for Computing Machinery.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multimodal Contrastive Representation Learning for Entity Alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584.

Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024. Improving Biomedical Entity Linking with Retrieval-enhanced Learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. *arXiv preprint arXiv:2308.11483*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging Large Language Models for Multiple Choice Question Answering. *arXiv preprint arXiv:2210.12353*.

Xiaoming Shi, Sendong Zhao, Yuxuan Wang, Xi Chen, Ziheng Zhang, Yefeng Zheng, and Wanxiang Che. 2022. Understanding Patient Query With Weak Supervision From Doctor Response. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2770–2777.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, 27.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-Domain Data Integration for Named Entity Disambiguation in Biomedical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4566–4575.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6397–6407.

Yuejia Xiang, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Zhenxi Lin, and Yefeng Zheng. 2021. OntoEA: Ontology-guided Entity Alignment via Joint Knowledge Graph Embedding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1117–1128.

Zhenran Xu, Yulin Chen, and Baotian Hu. 2023. Improving Biomedical Entity Linking with Cross-Entity Interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37.

Jinghui Yan, Yining Wang, Lu Xiang, Yu Zhou, and Chengqing Zong. 2020. A Knowledge-driven Generative Model for Multi-implication Chinese Medical Procedure Entity Normalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1490–1499.

Hongyi Yuan and Sheng Yu. 2021. Efficient Symptom Inquiring and Diagnosis via Adaptive Alignment of Reinforcement Learning and Classification. *arXiv preprint arXiv:2112.00733*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. BioBART: Pretraining and Evaluation of a Biomedical Generative Language Model. *arXiv preprint arXiv:2204.03905*.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. 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: Human Language Technologies*.

Yongqi Zhang, Quanming Yao, Ling Yue, Xian Wu, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. 2023. Emerging Drug Interaction Prediction Enabled by Flow-based Graph Neural Network with Biomedical Network. *Nature Computational Science*, 3(12):1023–1033.

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 Thirty-First International Joint Conference on Artificial Intelligence*.