Biomedical Entity Linking with Triple-aware Pre-Training

Xi Yan[1,*], Cedric Möller[1] and Ricardo Usbeck[2]

[1]Universität Hamburg, Edmund-Siemers-Allee 1, 20146, Hamburg, Germany
[2]Leuphana Universität Lüneburg , Universitätsallee 1, 21335 Lüneburg, Germany

Abstract

The large-scale analysis of scientific and technical documents is crucial for extracting structured knowledge from unstructured text. A key challenge in this process is linking biomedical entities, as these entities are sparsely distributed and often underrepresented in the training data of large language models (LLM). At the same time, those LLMs are not aware of high level semantic connection between different biomedical entities, which are useful in identifying similar concepts in different textual contexts. To cope with aforementioned problems, some recent works focused on injecting knowledge graph information into LLMs. However, former methods either ignore the relational knowledge of the entities or lead to catastrophic forgetting. Therefore, we propose a novel framework to pre-train the powerful generative LLM by a corpus synthesized from a KG. In the evaluations we are unable to confirm the benefit of including synonym, description or relational information. This work-in-progress highlights key challenges and invites further discussion on leveraging semantic information for LLm performance and on scientific document processing.

Keywords

Entity Linking, Scientific data, Deep learning, Semantic information

1. Introduction

Biomedical entity linking (EL) is a critical process in biomedical text mining that seeks to identify and associate relevant biological and medical entities mentioned in unstructured text with their corresponding identifiers in knowledge bases. EL systems have also been combined to promote the knowledge acquisition task[1]. Accurate recognition and linking of these entities are pivotal in promoting biomedical research, drug discovery, and personalized medicine [2]. Although substantial progress has been made in recent years, there is an ongoing need for refining methods and techniques employed for entity linking in the biomedical domain.

In this report, we present a novel approach that integrates linearized (in which a graph is traversed and encoded when producing the linearized representationHoyle et al. [3].) triples into the biomedical entity linking process while reevaluating the inclusion of synonym information. Our proposed method linearizes triples and considers them during the pre-training step. In past studies, synonym information, which involves using alternative names or terminologies for the same biomedical entity, has been proven to enhance entity linking when used during pre-training [4, 5]. Our study aims to build upon this existing knowledge by integrating both strategies and assessing their impact on performance.

Despite the reported benefits of synonym information in prior studies, our analysis of this approach, combined with the introduction of linearized triples [6], yielded different results. We find that incorporating linearized triples only lead to minimal improvements in our entity linking model's performance. Moreover, we are unable to confirm the purported advantages of including synonym information in our experiments, which stands in contrast to the findings of previous literature.

We highlight the limitations of our study and suggest possible avenues for future research to further advance biomedical entity linking techniques by building on our work with linearized triples and

Third International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data 2025
*Corresponding author.
xi.yan@uni-hamburg.de (X. Yan); cedric.moeller@uni-hamburg.de (C. Möller); ricardo.usbeck@leuphana.de (R. Usbeck)
https://www.hcds.uni-hamburg.de/hcds/head-hcds/xi-yan.html (X. Yan);
https://www.hcds.uni-hamburg.de/hcds/head-hcds/cedric-moeller.html (C. Möller);
https://www.leuphana.de/institute/iis/personen/ricardo-usbeck.html (R. Usbeck)
0009-0004-7379-4080 (X. Yan); 0000-0001-6700-3482 (C. Möller); 0000-0002-0191-7211 (R. Usbeck)

reevaluating synonym information. The code is available at our GitHub repo [1].

2. Related work

Entity Linking has a long history of research. Recent methods can be categorized into two types. First, discriminative methods that are based on the bi-encoder / cross-encoder pairing [7, 8, 9]. Both encoders are commonly BERT-like models. The bi-encoder encodes the description of each entity and matches it to the text by using an approximate nearest neighbor search. This is important as the next step, the cross-encoding, is expensive. Here, those neighbors are reranked by applying a cross-encoder to the concatenation of both, the input text and the entity description. The highest-ranked entity is then the final linked one. In the biomedical domain, the works by [10], [11], [12] and [13] fall into this category.

Another type of entity linker is based on generative models [14, 15, 5]. Here, instead of using some external description of an entity, the whole model memorizes the KG during training. The linked entity is then directly generated by the model. Such methods skip the problem of mining negatives which are crucial for a good performance of bi-encoder-based methods. BioLinkerAI [16] and Gallego et al [17] use the entity definitions and therasaurs (i.e., UMLS) to enhance the performance of LLM. Only the work by Yuan et al. [4] is based on such methods in the biomedical domain. As generative models lack the ability to incorporate external information, they alleviate this problem by introducing a pre-training stage where syntactical information from a knowledge graph is learned. This is especially important in the biomedical domain as entities often own a large variety of synonyms. We build upon their work by extending the pre-training regime to the inclusion of triple information.

3. Method

3.1. Task definition

Given are a text t, a set of marked mentions M_t in the text and a KG $\mathscr{G} = (\mathscr{E}, \mathscr{R}, E)$. The KG consists of a set of entities $\mathscr{E}$, a set of relations $\mathscr{R}$ and a set of edges composed of head entity, relation and tail entity $E \subseteq (\mathscr{E} \times \mathscr{R} \times \mathscr{E})$. The task is to identify the subset of entities $E_t \subseteq \mathscr{E}$ which the mentions M_t are referring to.

3.2. Model

In the vein of the work by [14], we model the problem as a sequence-to-sequence generation task. The input to the generative model is text and the output are the generated entity identifiers in the corresponding KGs. Similar to other works [14, 15, 4], we consider the definition of the concepts in the corresponding KGs as the unique textual representation of each concept. The definition and synonyms are short and unique, and will not introduce the problem of ambiguation of entities.

3.3. Pre-training

We linearize the information from synonym and triples in the pre-training stage. An overview of the pre-training and an example is give in Figure 1 . They are linearized into a synthesized corpora before feeding into the BART. We have tested 2 different settings for converting the triples, namely **line-by-line** and **all-in-one**. We add triple pre-training step on which add the triples information to the LLM, on top of synonym, which is used by [4].

In terms of the **synonym information**, we follow the setting by [4]. We first extract the description of the entity and convert it to a text of the following form:

$$\text{[BOS][ST] } s_e^a \text{ [ET] is defined as } c_e \text{ [EOS]} \tag{1}$$

[1]https://github.com/xixi019/bio-EL

Here, s_e^a stands for the synonym a and c_e for the description of entity e. This would be the input to the encoder of the generative model.

As an output the model has to generate:

$$\text{[BOS] } s_e^a \text{ is } s_e^b \text{ [EOS]} \tag{2}$$

This lets the model learn the connection between the different synonyms of the same entity.

Based on that, we introduce an additional pre-training step to incorporate more semantic information by utilising **triple information** from the underlying knowledge graph. A triple is of the form $< e, r, e' >$ which describes that a relationship r holds between entity e and e'. The input is here the same as for the synonym information. The output is of the form:

$$\text{[BOS] } s_e^a \; l_r \; s_{e'}^b \text{ [EOS]} \tag{3}$$

l_r is here the label of relation r. We denote this **line-by-line**. Furthermore, we experimented with an **all-in-one** pre-training approach of the form:

$$\text{[BOS] } s_e^a \; l_{r_1} \; s_{e_1}^{b_1} \; \ldots \; l_{r_n} \; s_{e_n}^{b_n} \text{ [EOS]} \tag{4}$$

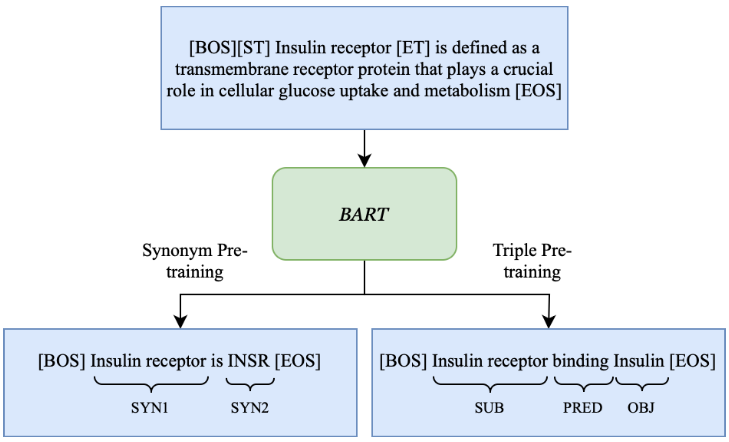

Figure 1: An overall workflow of our framework. We adopt different textualization formats for synonym information and triples. Both are included in the pre-training stage.

3.4. Fine-tuning

During fine-tuning, the model is trained for the actual entity linking task. The input to the generative model is the unlabelled biomedical text. To generate the linked entities, each mention is included in a template as follows:

$$\text{[BOS] } m_i \text{ is } s_e^a \text{ [EOS]} \tag{5}$$

The model then generates the entity identifier after the token "is". Similar to the work by Yuan et al. [4], we choose the synonym which is syntactically close to the corresponding mention in the text as the target entity identifier during fine-tuning.

The generated entity identifier is mapped back to the concrete entity in the final step via a lookup table. During inference, we restrict the possible output space by limiting it to the available entity names and synonyms. See Figure 2 for an overview of the pre-training with an example.

4. Evaluation

4.1. Pre-training Strategy

We use a synthesized corpus composed of triples, synonyms and descriptions from UMLS. More specifically, we decide to use a subset of UMLS, st21pv [18]. It is a well-connected KG with information

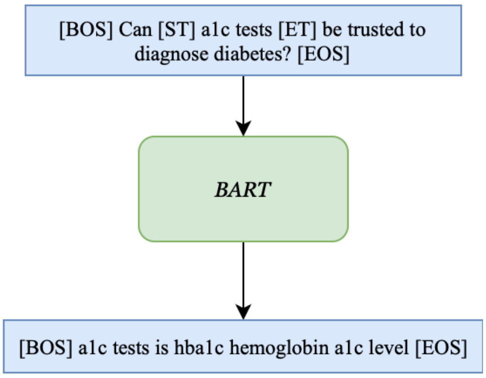

Figure 2: An overview of the fine-tuning stage

about concept definitions and synonyms. Specifically, 160K out of 2.37M concepts have definitions, 1.11M concepts have several synonyms and 68K concepts are connected to on average 8 triples as a subject in a single hop. During the pre-training step, we construct samples by iterating through each concept's synonyms and triples. Each concept is densely connected and the distribution of the number of triples a concept is connected to is skewed. For instance, some "popular" concepts are connected to over 1000 triples, while some are connected to only 1 triple. To avoid the class imbalance, we sample the included triples based on the relation frequencies.

To train the model with KG information, we linearize triples. Linearization refers to a special type of technique on converting graph to text, i.e., converting triples to one/more sentences which serve as input of the LLM.

We sample the included triples based on the relation frequencies. First, we gather the occurrence frequency of all relations in the KB by counting the number of triples this relation is connected to.

Both settings are trained under the same experiment setting with a batch size of 128. We save the best model within 12 training epochs. We experiment with BART-base, bioBART-Large, and bioBART-Base. We choose BART to align to the work of [4] so that we can make comparison about whether relational information is beneficial to the model. Note that we define the probability (P_r) of a relation r to be negatively related to the frequency. Then, for each concept in the KG, we collect its connected triples and segment the triples into different groups based on their relation r.

4.1.1. Fine-tuning

The model is fine-tuned on two established datasets, namely BC5CDR [19] and NCBI [20]. Those entity linking datasets are constructed on subsets of UMLS, making them perfect choices to test our model's performance on. Among the datasets, NCBI and BC5CDR are generated by annotating PubMed papers. On the other hand, NCBI and BC5CDR are annotated against Medical Subject Headings (MeSH) - a terminology knowledge graph for indexing and cataloging of biomedical information.

The statistics of the four datasets are exhibited in Table 1 below. As we can see, NCBI and BC5CDR (annotated on academic text) are smaller in size. Also NCBI and BC5CDR are dense in terms of the target entities they contain (14,967 and 268,162).

Table 1
Numbers of the samples in the training, development and test set

Nums	NCBI	BC5CDR
Train	5,784	9,285
Dev	787	9,515
Test	960	9,654
Entities	14,967	268,162

BART-large [21] is chosen as the generative model as it has been an established benchmark model for such tasks.

4.2. Results

We assess the performance of four distinct models in the entity linking task, including two of our own models, each pre-trained via either a line-by-line or all-in-one strategy, a synonym pre-trained model from [4] (denoted Syn-Only), and a basic BART model. We also include the recent papers which pretrains BART on biomedical domain [22] before finetuned on biomedical entity linking datasets and ResCNN [23] which achieves state-of-the-art results on various biomedical EL datasets. Each model undergoes fine-tuning specific to the entity linking task. Recall@1 for each model are presented in the Table 4.2. We limit ourselves to Recall@1 to follow the common practice when measuring entity linking performance without named entity recognition. The best-performing metrics are emphasized in bold.

Table 2
Recall@1 on BC5CDR and NCBI, which are PubMed articles annotated against MESH.

	BC5CDR	NCBI
Syn-Only	93.3%	91.9%
Syn-Only	92.68%	89.45%
All-in-one	92.86%	88.43%
Line-by-line	92.66%	90.00%
BART	92.58%	89.06%
BioBART-Large	93.01%	89.27%
BioBART-Base	**93.26**%	89.40%
ResCNN	91.7 %	**92.4**%

4.3. Analysis

Based on the table 4.2, our triple injection framework exceeds the BART baseline on the 2 benchmarks datasets. On BC5CDR and NCBI, the gain compared to BART is around 0.2% and 0.5%.

Does triple injection enhance model's capacity to link to the correct entity? The answer is yes, since over 2 datasets, the All-in-one or Line-by-line variants outperform the variant that was not trained on the linearized corpora for around 1% (Recall@1).

5. Conclusion

Our study seek to improve biomedical entity linking through the integration of linearized triples and synonym information. However, contrary to expectation, the incorporation of these elements leads to only minimal improvements in our EL model performance.

In conclusion, our study underscores the complexities of biomedical EL and prompts the need for more sophisticated approaches to improve its accuracy. A possible future extension of this work could be to explore more sophisticated methods to instruct the LLMs to learn external knowledge, such that the knowledge is injected in an efficient way which benefits the models in downstream tasks. For instance, by incorporating the KG information not just in a linearized manner but by exploiting the graph-structure with Graph Neural Networks [24], mutliple methods could be further developed.

6. Acknowledgments

This work has been partially supported by the Ministry of Research and Education within the project 'RESCUE-MATE: Dynamische Lageerstellung und Unterstützung für Rettungskräfte in komplexen

Krisensituationen mittels Datenfusion und intelligenten Drohnenschwärmen' (FKZ 13N16844), by the Federal Ministry for Economic Affairs and Climate Action of Germany in the project CoyPu (project number 01MK21007[G]). We utilized 2 x NVIDIA RTX A5000 24GB kindly provided by the NVIDIA Academic Hardware Grant Program. The authors have no competing interests to declare that are relevant to the content of this article.

Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-4 and Gramby in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] K. Noullet, A. Ourgani, M. Färber, A full-fledged framework for combining entity linking systems and components, in: Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 148–156. URL: https://doi.org/10.1145/3587259.3627556. doi:10.1145/3587259.3627556.

[2] P. Chandak, K. Huang, M. Zitnik, Building a knowledge graph to enable precision medicine, Scientific Data 10 (2023) 67. doi:10.1038/s41597-023-01960-3.

[3] A. M. Hoyle, A. Marasović, N. A. Smith, Promoting graph awareness in linearized graph-to-text generation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 944–956. URL: https://aclanthology.org/2021.findings-acl.82. doi:10.18653/v1/2021.findings-acl.82.

[4] H. Yuan, Z. Yuan, S. Yu, Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 4038–4048. doi:10.18653/v1/2022.naacl-main.296.

[5] Z. Xu, Y. Chen, B. Hu, Improving biomedical entity linking with cross-entity interaction, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 13869–13877. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26624. doi:10.1609/aaai.v37i11.26624.

[6] J. Li, T. Tang, W. X. Zhao, Z. Wei, N. J. Yuan, J.-R. Wen, Few-shot knowledge graph-to-text generation with pretrained language models, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 1558–1568. URL: https://aclanthology.org/2021.findings-acl.136. doi:10.18653/v1/2021.findings-acl.136.

[7] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 6397–6407. doi:10.18653/v1/2020.emnlp-main.519.

[8] L. Logeswaran, M. Chang, K. Lee, K. Toutanova, J. Devlin, H. Lee, Zero-shot entity linking by reading entity descriptions, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 3449–3460. doi:10.18653/v1/p19-1335.

[9] T. Ayoola, S. Tyagi, J. Fisher, C. Christodoulopoulos, A. Pierleoni, Refined: An efficient zero-shot-capable approach to end-to-end entity linking, in: A. Loukina, R. Gangadharaiah, B. Min (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2022, Hybrid: Seattle, Washington, USA + Online, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 209–220. doi:10.18653/v1/2022.naacl-industry.24.

[10] R. Angell, N. Monath, S. Mohan, N. Yadav, A. McCallum, Clustering-based inference for biomedical entity linking, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 2598–2608. doi:10.18653/v1/2021.naacl-main.205.

[11] M. Varma, L. J. Orr, S. Wu, M. Leszczynski, X. Ling, C. Ré, Cross-domain data integration for named entity disambiguation in biomedical text, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 4566–4575. doi:10.18653/v1/2021.findings-emnlp.388.

[12] D. Agarwal, R. Angell, N. Monath, A. McCallum, Entity linking and discovery via arborescence-based supervised clustering, CoRR abs/2109.01242 (2021). URL: https://arxiv.org/abs/2109.01242. arXiv:2109.01242.

[13] R. Bhowmik, K. Stratos, G. de Melo, Fast and effective biomedical entity linking using a dual encoder, in: E. Holderness, A. Jimeno-Yepes, A. Lavelli, A. Minard, J. Pustejovsky, F. Rinaldi (Eds.), Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, LOUHI@EACL, Online, April 19, 2021, Association for Computational Linguistics, 2021, pp. 28–37. URL: https://www.aclweb.org/anthology/2021.louhi-1.4/.

[14] N. D. Cao, G. Izacard, S. Riedel, F. Petroni, Autoregressive entity retrieval, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.

[15] N. D. Cao, L. Wu, K. Popat, M. Artetxe, N. Goyal, M. Plekhanov, L. Zettlemoyer, N. Cancedda, S. Riedel, F. Petroni, Multilingual autoregressive entity linking, Trans. Assoc. Comput. Linguistics 10 (2022) 274–290. URL: https://doi.org/10.1162/tacl_a_00460. doi:10.1162/tacl_{a}{_{0}{0}{4}}{6}0.

[16] A. Sakor, K. Singh, M.-E. Vidal, Biolinkerai: Capturing knowledge using llms to enhance biomedical entity linking, in: International Conference on Web Information Systems Engineering, Springer, 2024, pp. 262–272.

[17] F. Gallego, P. Ruas, F. M. Couto, F. J. Veredas, Enhancing cross-encoders using knowledge graph hierarchy for medical entity linking in zero-and few-shot scenarios, Knowledge-Based Systems 314 (2025) 113211.

[18] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with UMLS concepts, in: 1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019, 2019. URL: https://doi.org/10.24432/C5G59C. doi:10.24432/C5G59C.

[19] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, Biocreative V CDR task corpus: a resource for chemical disease relation extraction, Database J. Biol. Databases Curation 2016 (2016). URL: https://doi.org/10.1093/database/baw068. doi:10.1093/database/baw068.

[20] R. I. Dogan, R. Leaman, Z. Lu, NCBI disease corpus: A resource for disease name recognition and concept normalization, J. Biomed. Informatics 47 (2014) 1–10. URL: https://doi.org/10.1016/j.jbi.2013.12.006. doi:10.1016/j.jbi.2013.12.006.

[21] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

[22] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, BioBART: Pretraining and evaluation of a biomedical generative language model, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 97–109. URL: https://aclanthology.org/2022.bionlp-1.9.

[23] T. Lai, H. Ji, C. Zhai, BERT might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 1631–1639. URL: https://aclanthology.org/2021.findings-emnlp.140. doi:10.18653/v1/2021.findings-emnlp.140.

[24] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S. Y. Philip, A comprehensive survey on graph neural networks, IEEE transactions on neural networks and learning systems 32 (2020) 4–24.