Published in final edited form as:
Int ACM SIGIR Conf Res Dev Inf Retr. 2023 July ; 2023: 2052–2056. doi:10.1145/3539618.3591997.

HiPrompt: Few-Shot Biomedical Knowledge Fusion via Hierarchy-Oriented Prompting

Jiaying Lu,
Emory University, USA

Jiaming Shen,
Google Research, USA

Bo Xiong,
University of Stuttgart, Germany

Wenjing Ma,
Emory University, USA

Steffen Staab,
University of Stuttgart, Germany, University of Southampton, UK

Carl Yang
Emory University, USA

Abstract

Medical decision-making processes can be enhanced by comprehensive biomedical knowledge bases, which require fusing knowledge graphs constructed from different sources via a uniform index system. The index system often organizes biomedical terms in a hierarchy to provide the aligned entities with fine-grained granularity. To address the challenge of scarce supervision in the biomedical knowledge fusion (BKF) task, researchers have proposed various unsupervised methods. However, these methods heavily rely on ad-hoc lexical and structural matching algorithms, which fail to capture the rich semantics conveyed by biomedical entities and terms. Recently, neural embedding models have proved effective in semantic-rich tasks, but they rely on sufficient labeled data to be adequately trained. To bridge the gap between the scarce-labeled BKF and neural embedding models, we propose HiPrompt, a supervision-efficient knowledge fusion framework that elicits the few-shot reasoning ability of large language models through hierarchy-oriented prompts. Empirical results on the collected KG-Hi-BKF benchmark datasets demonstrate the effectiveness of HiPrompt.

Keywords

Biomedical Knowledge Fusion; Few-Shot Prompting; Large Language Models for Resource-Constrained Field; Retrieve; Re-Rank

jiaying.lu@emory.edu .

1 INTRODUCTION

In the biomedical field, there exists a lot of knowledge acquired from clinical practice guidelines, medical records, and publications, accumulated from different research laboratories and healthcare institutions [8, 34, 36]. Recently, knowledge graphs (KGs) have emerged as a compelling technique to efficiently represent, organize, and distribute knowledge. A biomedical KG stores the properties of biomedical entities and their relations. Researchers' constant endeavors in manually curating biomedical KGs have led to the existence of many domain-specific and application-oriented KGs. However, these well-annotated biomedical KGs are scattered in various data formats, which hinders their off-the-shelf usability.

Fusing KGs from multiple sources into an accurate and comprehensive knowledge base can greatly support clinical decision-making [13, 28]. A common practice is to align entities of KGs with standard hierarchical index systems (*i.e. biomedical hierarchies*) [4, 14, 30, 44]. The hierarchy allows entities to be aligned and analyzed more precisely with fine-grained granularity, which is beneficial to many downstream tasks [21, 31, 32, 40, 43]. Moreover, the biomedical hierarchy is well maintained with periodic upgrades to incorporate newly emerging biomedical terms, thus enabling scalable integration with multiple KGs. In this work, we study the biomedical knowledge fusion (*BKF*) problem that aims to align entities from biomedical KGs into terms from the biomedical hierarchy. Figure 1 gives a toy example of the BKF task. The BKF task is challenging due to the following characteristics. First, inconsistent naming vocabularies are used in different resources, as they are developed independently by different groups of specialists. Second, unlike the existing KG entity alignment problem [38, 47] that contains many labeled entity-entity pairs as training samples, biomedical knowledge integration is supervision-scarce. Third, the topology of a KG and a hierarchy are very different, where the KG is a general graph, while the hierarchy is a directed acyclic graph.

Existing research.

Pioneer studies on BKF mainly rely on the biomedical thesaurus to normalize words and match lexical to establish alignment between KGs and the hierarchy [13, 24, 28, 36]. Later, researchers explore combing first-order logic [15], probabilistic alignment [37], or non-literal string comparisons [11] with lexical matching for unsupervised BKF. However, these methods fail to capture the rich semantics conveyed in entities and terms (*e.g.*, synonyms, definitions, types), which are essential to handle the inconsistent naming conventions from multi-sources. Another line of work leverage neural embedding models [9, 19, 20, 38, 46] to represent entities as dense vectors using semantic attributes, structural properties, and alignment supervisions. These models perform better than unsupervised models when sufficient training samples are available. However, the scarcity of supervision in the BKF problem leads to the underfitting of these data-eager neural models. Moreover, none of the existing methods explicitly leverages the hierarchical structure of terms in the biomedical hierarchy.

Present work.

To address above challenges, we present **HiPrompt**, a few-shot BKF framework via **Hi**erarchy-Oriented **Promp**ting. HiPrompt employs a large language model (LLM) to generatively propose terms from the hierarchy to be aligned with entities from the KG. The key insight is that LLMs [7, 10, 39, 48] can be rapidly adapted to an unseen task via the gradient-free "prompt-based learning" [35, 41], thus removing the dependencies on the task-specific supervision. HiPrompt applies prompt-based learning with a curated task description for the BKF task and a tiny number of demonstrations generated from the few-shot samples. This mimics the procedure of how humans accomplish a new task by learning from previous experiences and generalizing them to a new context. Moreover, we add the hierarchical context to the prompts to further improve the performance of HiPrompt. To evaluate the performance of our proposed HiPrompt, we create *KG-Hi-BKF*, a new benchmark for BKF with two datasets collected from two biomedical KGs [6, 50] and one disease hierarchy [30] with manual verification. Empirical results demonstrate the effectiveness of our HiPrompt framework, which largely outperforms both conventional unsupervised lexical matching models and neural semantic embedding models.

2 BIOMEDICAL KNOWLEDGE FUSION

2.1 Problem Definition

BKF aims at aligning existing specialized biomedical KGs into a uniform biomedical index system that can be represented by a hierarchy. We define the biomedical KG and hierarchy as follows: A biomedical KG is a multi-relation graph $\mathcal{G} = (E, R, RT)$, where E, R, RT are a set of various types of entities, a set of relation names, and $RT \in E \times R \times E$ is the set of relational triples, respectively. A biomedical hierarchy is a directed acyclic graph (DAG) $\mathcal{H} = (T, TP)$, where T is a set of terms, and $TP \in T \times T$ is a set of Hhypernym-hyponymy term pairs, respectively. The topology differences between KG and hierarchy distinguish our BKF task from other related tasks (*e.g.*, entity alignment, KG integration). Moreover, both entities E and terms T contain rich associated semantic attributes (*e.g.*, definition, synonyms). Finally, we define our task as follows:

Definition 2.1 (biomedical knowledge fusion). Given a biomedical KG $\mathcal{G}$, a biomedical hierarchy $\mathcal{H}$, a set of pre-aligned entity-term pairs $[e_a, t_a]_{a=1}^{M}$, and a set of unaligned entities $[e_1, e_2, \cdot, e_N] \in \mathcal{G}$. The goal is to link each unaligned entity to the hierarchy $LK = \{(e_i, t_j) | e_i \in \mathcal{G}, t_j \in \mathcal{H}\}$ such that t_j is the most specific term in the hierarchy for entity e_i in KG. In our work, we focus on the few-shot settings where the sample size M is very small to reflect the scarcity of labeled data that is ubiquitous in the biomedical field.

2.2 Technical Details of HiPrompt

Figure 2 shows the overall architecture of our proposed HiPrompt framework. To tackle the BKF task with limited training samples, our key insight is to utilize LLMs via hierarchy-oriented prompting. However, LLMs can not accommodate very lengthy input prompts (*e.g.*, GPT-3 only supports up to 4096 tokens) that contain all candidate terms along with their hierarchy contexts. A feasible workaround is to exhaustively examine each candidate term given the query entity, but the inference cost would be dramatic [23]. Therefore, we propose to use the *retrieve and re-rank* [12, 22, 42] approach to resolve the above challenges.

Retrieval Module.—The retriever provides an efficient solution for coarse-grained candidate filtering, thus reducing the overall inference cost of HiPrompt. Given one entity query e_i from the KG $\mathscr{G}$ and all candidate terms T from the hierarchy $\mathscr{H}$, the retriever produces a coarsely ranked candidate list $(t_1', t_2', \cdots, t_K')$, to avoid unnecessary computations for the LLM-based re-ranker. HiPrompt framework is flexible so that any unsupervised ranking function (*e.g.*, TF-IDF [27], LDA [3]) can be used to generate the ranked list. In practice, we choose the unsupervised BM25 [26] as the ranking function. Since entities and concepts have rich attributive and structural information, we further utilize these two types of information to expand [2] query entities and candidate terms.

Re-Ranking Module.—Given the query entity e_i and the coarsely ranked candidate list $(t_1', t_2', \cdots, t_K')$, we request the LLM to rerank the list to $(t_1, t_2, \ldots t_K)$ where t_1 is the most specific term of e_i via the gradient-free prompt-based learning. Figure 2 provides an example of the input prompt and the response of the re-ranker. The input prompt is composed of (1) curated textual ***task description***, (2) illustrative ***demonstration*** from few-show samples, and (3) the ***test prompt*** constructed from the query entity and the coarsely ranked list. The LLM-based re-ranker essentially tackles the BKF task by estimating the conditional probability: $P_{LLM}(w_1, w_2, \ldots, w_n | prompt)$, where $(w_1, \ldots, w_n)$ is the output word sequence with variable lengths. The desired re-ranked list can be converted from the output sequence by a simple mapping function $(t_1, t_2, \ldots, t_K) = f(w_1, w_2, \ldots, w_{n.})$.

For the template of demonstration, we use the query entity to form the question string "Query: $\{e_i\}$", the coarse candidate list to form the choice string "Choices: $\{t_1'; t_2'; \ldots t_K'\}$", and the ground truth to form the answer string "Answer: $\{t_1; t_2; \ldots, t_K\}$". While there is no such ground truth sample in the zero-shot setting, we propose the *pseudo demonstration* technique which adopts out-of-domain entity-term pairs to showcase what is the perspective format. Both real and pseudo demonstrations are essential to generate output sequences in the consistent format [16, 29]. For the test prompt, we use the same template of the demonstration, while leaving the answer string as "Answer:" for LLM to predict what comes next. To further elicit LLMs with hierarchical constraints and dependencies of candidate terms, we propose the novel ***test prompt with hierarchy context*** where hypernyms of each candidate term are included in the context string. More specifically, we traverse the biomedical hierarchy $\mathscr{T}$ to locate the hypernym terms $t_{i,p_1}', \cdots, t_{i,p_j}'$ of a candidate term t_i'. Therefore, the context string is formed as "Contexts: $\{t_1'$ isA $t_{1,p}'; \ldots; t_K'$ isA $t_{K,p}'\}$".

3 EXPERIMENTS

Benchmark Datasets.

We use the following data sources to create our KG-Hi-BKF benchmark[1]: (1) SDKG [50]: a disease-centric KG that covers five cancers and six non-cancer diseases. (2) repoDB [6]:

[1] KG-Hi-BKF benchmark is available at https://doi.org/10.6084/m9.figshare.21950282.

we adopt their original triples, and generate entity attributes by querying DrugBank [44] and UMLS Metathesaurus [4]. (3) DzHi [30]: a hierarchy derived from the widely used Disease Ontology [30] which has a depth of 13. We first use the mapping existing in the resources themselves, which leads to many-to-many linkages between two KBs. We further manually verify the correctness of the many-to-many linkages and curate the datasets to the correct stage. Table 2 shows the statistics of the created benchmark. As can be seen, the linkages follow the one-to-one assumption [38], and the scale of labeled entity-term pairs is very small.

Compared Models.

We compare HiPrompt to the following two sets of baselines: (a) *Non-neural conventional models*: (a.1) **Edit Dist** [25] that quantifies the distance between entities and terms by the edit distance of their names. (a.2) **BM25** [26] that ranks a set of documents based on the query tokens appearing in each document. (a.3) **LogMap** [15] that matches entities and terms via logical constraints and semantical features. (a.4) **PARIS** [37] that provides a off-the-shelf fusion tool empowered by a parameter tuning-free probabilistic model. (a.5) **AML** [11] that is based on non-literal string comparison algorithms. is a probabilistic matching system based on probability estimates. (b) *Neural embedding models*: (b.1) **SapBERT** [17] that learns to self-align synonymous biomedical entities through a Transformer. (b.2) **MTransE** [9] that extends the translational KG embedding method TransE [5] to multi-language system entity alignment by axis calibration and linear transformations. (b.3) **SelfKG** [18] that designs a self-negative sampling strategy to push sampled negative pairs far away from each other when no labeled positive pairs are available.

Quantitative evaluations.

We mainly focus on zero-shot and one-shot settings, and utilize the remaining labeled samples as the test set to report quantitative results. Several *strict* and *lenient* evaluation metrics are used. For strict metrics that appreciate only the exact correct prediction, we adopt **Hits@k** and mean reciprocal rank (**MRR**). For lenient metrics that also reward near-hits, we adopt **nDCG@k** with exponential decay [1] and hierarchy-based term relatedness score **WuP** [45]. All compared baselines are executed with their recommended hyperparameters. For all non-neural conventional models, we only report the zero-shot results as they are unsupervised methods. For neural embedding methods, we report the zero-shot results utilizing released model weights (SapBERT) or conducting self-supervised training (SelfKG), while reporting the one-shot results by fine-tuning these models (SapBERT, MTransE) on the one demonstrative training sample. For our HiPrompt, we use GPT-3 [7] as the LLM for re-ranker and set its temperature hyperparameters as 0 to lower the completion randomness. Using a single prompt template is sufficient since initial exploration shows that various templates do not have a significant impact on model performance. We exclude the use of automatic prompt generation techniques [33, 49] due to the limited availability of training data.

Main Results.

Table 1 shows the quantitative results for zero-shot and one-shot settings. HiPrompt largely outperforms all other methods in all evaluation metrics under both settings, which demonstrates the effectiveness of the proposed hierarchy-oriented prompting. Under the zero-shot setting, the non-neural unsupervised baseline LogMap achieves the second-best performance. All examined models can successfully generate predictions except AML throws out-of-memory (OOM) errors on the SDKG-DzHi dataset. PARIS performs worst in the zero-shot setting because it can not predict aligned terms for each query entity. Instead, PARIS produces the alignment based on its own ad-hoc threshold. MTransE performs worst in the one-shot setting since it is underfitting using just one training sample. Comparing the same models (SapBERT, HiPrompt) between zero-shot and one-shot settings, we observe the performance differences are negligible, thus indicating that effectively eliciting the adaptive reasoning ability is one of the key factors to tackling supervision-scarce BKF problem.

Ablation Studies.

We further conduct ablation studies to evaluate the impact of our hierarchy-oriented techniques. Table 3 compares the different expansion strategies for HiPrompt's retrieval module. As can be seen, if expanding the KG entities and hierarchy terms with both attributive and structural features ("*+Atr.+Str.*" variant), the retriever can achieve the best Hits@K performance. Table 4 compares different LLMs and different prompts for HiPrompt's reranking module. Among the examined LLMs, GPT-3 with 175 billion parameters surpasses GPT-JT [39] with 6B parameters and OPT-6.7B [48] with 6.7B parameters due to its large parameter space. When adding the proposed hierarchy context to the name-only prompts, every LLM achieves better performance on all metrics, thus demonstrating the importance of explicit hierarchy-oriented information. We also observe that improvements for GPT-JT and OPT-6.7B are more significant than GPT-3, since GPT-3 may already have such hierarchical information encoded.

Case Studies.

Figure 3 shows the fusion results from BM25, EditDist, and HiPrompt. In general, HiPrompt can find the most specific terms in the hierarchy for the query entities, by satisfying the semantic similarities and hierarchical constraints simultaneously. For instance, HiPrompt recognizes that "*immune system disease*" is the most appropriate for the query "*immune suppression*", rather than its hypernym "*disease of anatomical entity*" that is too general, or hyponyms such as "*immune system cancer*" or "*allergic disease*" that are too specific. On the other hand, EditDist only considers lexical matching, thereby ignoring the different naming conventions of the same biomedical concepts. BM25 also mainly relies on lexical matching, but it incorporates the names, definitions, and synonyms of biomedical terms during the matching, resulting in better performance in handling various names. However, BM25 ignores the hierarchical information, which leads to the inappropriate granularity of aligned terms (*e.g.*, the term "*epidemic typhus*" is too broad for the query entity "*typhus, epidemic Louse-Borne*").

4 CONCLUSIONS

This paper studies how to automatically fuse KGs into a standard hierarchical index system with scarce labeled data. Our novel framework, HiPrompt, uses hierarchy-oriented prompts to elicit the few-shot reasoning ability of large language models and is designed to be supervision-efficient. Performance comparison on the newly collected KG-Hi-BKF benchmark with two datasets demonstrates the effectiveness of HiPrompt. Interesting future directions for BKF include: (1) exploring an automatic way to generate hierarchy-aware prompts to further reduce manual intervention; (2) expanding the scope of biomedical knowledge fusion to allow the hierarchy to dynamically grow with the aligned entities.

REFERENCES

[1]. Balog Krisztian and Neumayer Robert. 2012. Hierarchical target type identification for entity-oriented queries. In CIKM.

[2]. Billerbeck Bodo and Zobel Justin. 2005. Document expansion versus query expansion for ad-hoc retrieval. In Proceedings of the 10th Australasian Document Computing Symposium.

[3]. Blei David M, Ng Andrew Y, and Jordan Michael I. 2003. Latent Dirichlet Allocation. JMLR (2003).

[4]. Bodenreider Olivier. 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research (2004).

[5]. Bordes Antoine, Usunier Nicolas, García-Durán Alberto, Weston Jason, and Yakhnenko Ok-sana. 2013. Translating Embeddings for Modeling Multi-relational Data. In NeurIPS.

[6]. Brown Adam S and Patel Chirag J. 2017. A standard database for drug repositioning. Scientific data (2017).

[7]. Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Jared D Kaplan Prafulla Dhariwal, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, et al. 2020. Language models are few-shot learners. NeurIPS (2020).

[8]. Chandak Payal, Huang Kexin, and Zitnik Marinka. 2022. Building a knowledge graph to enable precision medicine. bioRxiv (2022).

[9]. Chen Muhao, Tian Yingtao, Yang Mohan, and Zaniolo Carlo. 2017. Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment. In IJCAI.

[10]. Dong Xiangjue, Lu Jiaying, Wang Jianling, and Caverlee James. 2023. Closed-book Question Generation via Contrastive Learning. In EACL.

[11]. Faria Daniel, Pesquita Catia, Santos Emanuel, Palmonari Matteo, Cruz Isabel F, and Couto Francisco M. 2013. The agreementmakerlight ontology matching system. In ODBASE.

[12]. Glass Michael, Rossiello Gaetano, Chowdhury Md Faisal Mahbub, Naik Ankita, Cai Pengshan, and Gliozzo Alfio. 2022. Re2G: Retrieve, Rerank, Generate. In NAACL.

[13]. Himmelstein Daniel Scott, Lizee Antoine, Hessler Christine, Brueggeman Leo, Chen Sabrina L, Hadley Dexter, Green Ari, Khankhanian Pouya, and Baranzini Sergio E. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife (2017).

[14]. Jiang Shuai, Qian Qiheng, Zhu Tongtong, Zong Wenting, Shang Yunfei, Jin Tong, Zhang Yuansheng, Chen Ming, Wu Zishan, Chu Yuan, et al. 2023. Cell Taxonomy: a curated repository of cell types with multifaceted characterization. Nucleic Acids Research (2023).

[15]. Jiménez-Ruiz Ernesto and Grau Bernardo Cuenca. 2011. Logmap: Logic-based and scalable ontology matching. In ISWC.

[16]. Kojima Takeshi, Gu Shixiang Shane, Reid Machel, Matsuo Yutaka, and Iwasawa Yusuke. 2022. Large Language Models are Zero-Shot Reasoners. In ICML 2022 Workshop on Knowledge Retrieval and Language Models.

[17]. Liu Fangyu, Shareghi Ehsan, Meng Zaiqiao, Basaldella Marco, and Collier Nigel. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In NAACL.

[18]. Liu Xiao, Hong Haoyun, Wang Xinghao, Chen Zeyi, Kharlamov Evgeny, Dong Yuxiao, and Tang Jie. 2022. SelfKG: Self-Supervised Entity Alignment in Knowledge Graphs. In The Web Conference.

[19]. Lu Jiaying, Dong Xiangjue, and Yang Carl. 2023. Weakly Supervised Concept Map Generation through Task-Guided Graph Translation. IEEE Transactions on Knowledge and Data Engineering (2023).

[20]. Lu Jiaying and Yang Carl. 2022. Open-World Taxonomy and Knowledge Graph Co-Learning. In 4th Conference on Automated Knowledge Base Construction.

[21]. Ma Wenjing, Lu Jiaying, and Wu Hao. 2023. Cellcano: supervised cell type identification for single cell ATAC-seq data. Nature Communications (2023).

[22]. Matsubara Yoshitomo, Vu Thuy, and Moschitti Alessandro. 2020. Reranking for efficient transformer-based answer selection. In SIGIR.

[23]. Patterson David, Gonzalez Joseph, Le Quoc, Liang Chen, Munguia Lluis-Miquel, Rothchild Daniel, So David, Texier Maud, and Dean Jeff. 2021. Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350 (2021).

[24]. Ren Xiang, Shen Jiaming, Qu Meng, Wang Xuan, Wu Zeqiu, Zhu Qi, Jiang Meng, Tao Fangbo, Sinha Saurabh, Liem David, Ping Peipei, Weinshilboum Richard M., and Han Jiawei. 2017. Life-iNet: A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences. In ACL.

[25]. Ristad Eric Sven and Yianilos Peter N. 1998. Learning string-edit distance. TPAMI (1998).

[26]. Robertson Stephen, Zaragoza Hugo, et al. 2009. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval (2009).

[27]. Salton Gerard and Buckley Chris. 1988. Term-weighting approaches in automatic text retrieval. Information processing & management (1988).

[28]. Santos Alberto, Colaço Ana R, Nielsen Annelaura B, Niu Lili, Strauss Maximilian, Geyer Philipp E, Coscia Fabian, Wewer Albrechtsen Nicolai J, Mundt Filip, Jensen Lars Juhl, and Mann Matthias. 2022. A knowledge graph to interpret clinical proteomics data. Nature Biotechnology (2022).

[29]. Schick Timo and Schütze Hinrich. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In EACL.

[30]. Schriml Lynn M, Munro James B, Schor Mike, Olley Dustin, McCracken Carrie, Felix Victor, Baron J Allen, Jackson Rebecca, Bello Susan M, Bearer Cynthia, et al. 2022. The human disease ontology 2022 update. Nucleic acids research (2022).

[31]. Shen Jiaming and Han Jiawei. 2022. Automated Taxonomy Discovery and Exploration. Springer Nature.

[32]. Shen Jiaming, Qiu Wenda, Meng Yu, Shang Jingbo, Ren Xiang, and Han Jiawei. 2021. TaxoClass: Hierarchical multi-label text classification using only class names. In NAACL.

[33]. Shin Taylor, Razeghi Yasaman, Logan Robert L IV, Wallace Eric, and Singh Sameer. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In EMNLP.

[34]. Sigdel Dibakar, Kyi Vincent, Zhang Aiden, Shaun P Setty David Liem, Shi Yu, Wang Xuan, Shen Jiaming, Wang Wei, Han Jiawei, and Ping Peipei. 2019. Cloud-Based Phrase Mining and Analysis of User-Defined Phrase-Category Association in Biomedical Publications. JoVE 144 (2019).

[35]. Singhal Karan, Azizi Shekoofeh, Tu Tao, Mahdavi S Sara, Wei Jason, Chung Hyung Won, Scales Nathan, Tanwani Ajay, Cole-Lewis Heather, Pfohl Stephen, et al. 2022. Large Language Models Encode Clinical Knowledge. arXiv preprint arXiv:2212.13138 (2022).

[36]. Su Chang, Hou Yu, Rajendran Suraj, Maasch Jacqueline RMA, Abedi Zehra, Zhang Haotan, Bai Zilong, Cuturrufo Anthony, Guo Winston, Chaudhry Fayzan F, et al. 2021. Biomedical Discovery through the integrative Biomedical Knowledge Hub (iBKH). medRxiv (2021).

[37]. Suchanek Fabian M, Abiteboul Serge, and Senellart Pierre. 2011. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. VLDB (2011).

[38]. Sun Zequn, Zhang Qingheng, Hu Wei, Wang Chengming, Chen Muhao, Akrami Farahnaz, and Li Chengkai. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. VLDB (2020).

[39]. Together. 2023. GPT-JT-6B. https://huggingface.co/togethercomputer/GPT-JT-6B-v1. Accessed on February 14, 2023.

[40]. Tsatsaronis George, Balikas Georgios, Malakasiotis Prodromos, Partalas Ioannis, Zschunke Matthias, Michael R Alvers Dirk Weissenborn, Krithara Anastasia, Petridis Sergios, Polychronopoulos Dimitris, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics (2015).

[41]. Wang Boshi, Min Sewon, Deng Xiang, Shen Jiaming, Wu You, Zettlemoyer Luke, and Sun Huan. 2022. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. arXiv preprint arXiv:2212.10001 (2022).

[42]. Wang Lidan, Lin Jimmy, and Metzler Donald. 2011. A cascade ranking model for efficient ranked retrieval. In SIGIR.

[43]. Wang Lu, Tang Ruiming, He Xiaofeng, and He Xiuqiang. 2022. Hierarchical imitation learning via subgoal representation learning for dynamic treatment recommendation. In WSDM.

[44]. Wishart David S, Feunang Yannick D, Guo An C, Lo Elvis J, Marcu Ana, Grant Jason R, Sajed Tanvir, Johnson Daniel, Li Carin, Sayeeda Zinat, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic acids research (2018).

[45]. Wu Zhibiao and Palmer Martha. 1994. Verbs semantics and lexical selection. In ACL.

[46]. Xiong Bo, Potyka Nico, Tran Trung-Kien, Nayyeri Mojtaba, and Staab Steffen. 2022. Faithful Embeddings for EL++ Knowledge Bases. In ISWC.

[47]. Xu Chengjin, Su Fenglong, Xiong Bo, and Lehmann Jens. 2022. Time-aware Entity Alignment using Temporal Relational Attention. In WWW.

[48]. Zhang Susan, Roller Stephen, Goyal Naman, Artetxe Mikel, Chen Moya, Chen Shuohui, Dewan Christopher, Diab Mona, Li Xian, Lin Xi Victoria, et al. 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022).

[49]. Zhang Zhuosheng, Zhang Aston, Li Mu, and Smola Alex. 2023. Automatic Chain of Thought Prompting in Large Language Models. In The Eleventh International Conference on Learning Representations.

[50]. Zhu Chaoyu, Yang Zhihao, Xia Xiaoqiong, Li Nan, Zhong Fan, and Liu Lei. 2022. Multimodal reasoning based on knowledge graph embedding for specific diseases. Bioinformatics (2022).

CCS CONCEPTS

• **Applied computing** → **Health care information systems**; • **Information systems** → *Retrieval models and ranking*.

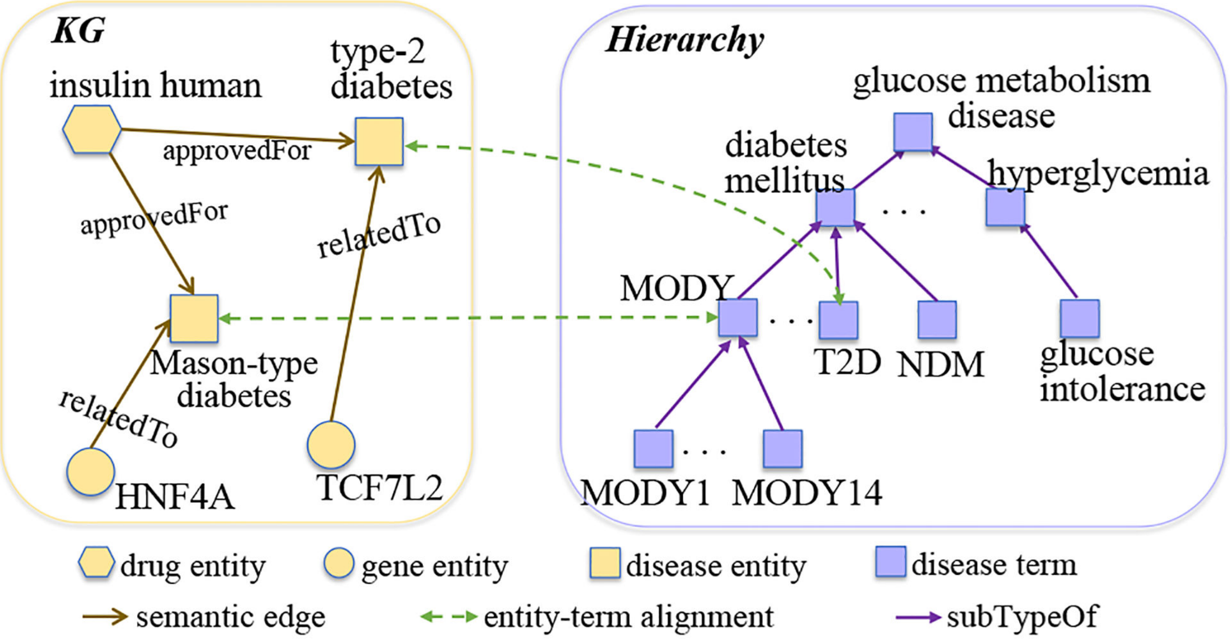
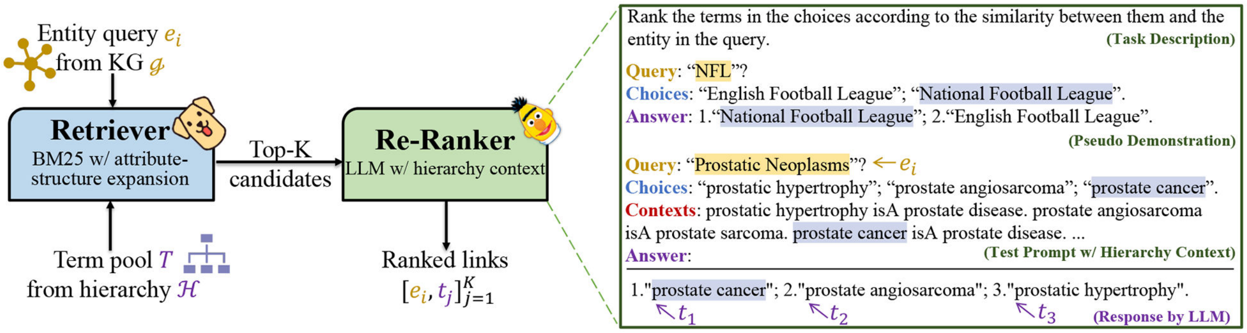


Figure 1:

A toy example of BKF to find entity-term alignment between KG and hierarchy. *Left*: A KG containing biomedical entities. *right*: A hierarchy containing biomedical terms.



Rank the terms in the choices according to the similarity between them and the entity in the query. **(Task Description)**

Query: "NFL"?
Choices: "English Football League"; "National Football League".
Answer: 1."National Football League"; 2."English Football League".
(Pseudo Demonstration)

Query: "Prostatic Neoplasms"? ← e_i
Choices: "prostatic hypertrophy"; "prostate angiosarcoma"; "prostate cancer".
Contexts: prostatic hypertrophy isA prostate disease. prostate angiosarcoma isA prostate sarcoma. prostate cancer isA prostate disease. ...
Answer: **(Test Prompt w/ Hierarchy Context)**

1."prostate cancer"; 2."prostate angiosarcoma"; 3."prostatic hypertrophy".
↖t_1 ↖t_2 ↖t_3 **(Response by LLM)**

Figure 2:

Overview of our HiPrompt framework, with a zoom-in on the LLM-based re-ranker.

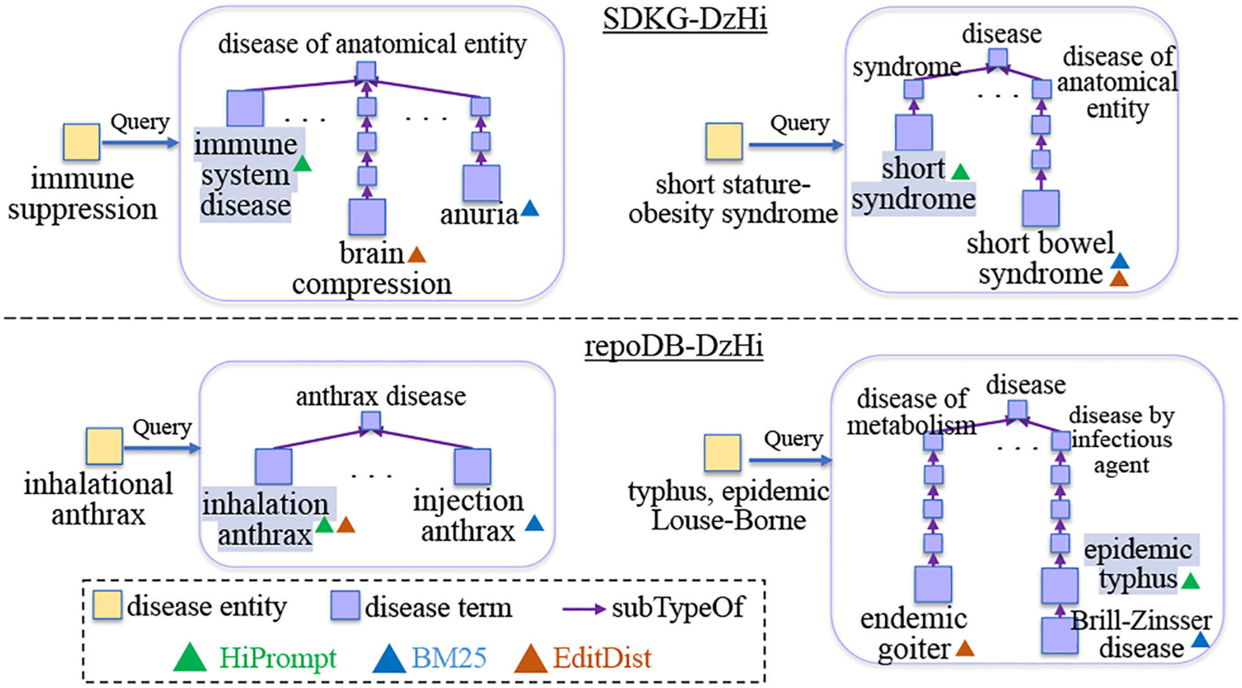


Figure 3:

Case Studies on unlabeled data. Terms highlighted in violet denote the correct alignments for query entities.

Table 1:

Main experiment results (in percentages).

Setting	Model	SDKG-DzHi						repoDB-DzHi					
		Hits@1	**Hits@3**	**nDCG@1**	**nDCG@3**	**WuP**	**MRR**	**Hits@1**	**Hits@3**	**nDCG@1**	**nDCG@3**	**WuP**	**MRR**
Zero-shot	Edit Dist	65.51	70.39	68.08	50.82	85.53	68.69	68.69	71.37	71.71	54.15	85.21	70.71
	BM25	73.07	87.40	77.56	63.01	91.97	81.06	59.38	74.75	70.33	64.51	90.71	68.84
	LogMap	75.75	79.06	76.97	54.82	85.06	77.38	86.60	87.73	87.38	60.79	91.68	87.09
	PARIS	22.68	22.68	23.15	16.13	43.85	22.68	6.35	6.35	6.42	4.44	32.28	6.35
	AML	OOM	OOM	OOM	OOM	OOM	OOM	78.00	78.56	78.67	54.90	86.02	78.26
	SapBERT	69.61	87.24	76.38	63.86	93.78	78.97	75.04	90.69	81.24	73.51	94.25	83.61
	SelfKG	57.95	69.45	58.98	47.29	74.25	64.70	72.78	81.10	75.95	63.78	88.41	77.71
	HiPrompt	**90.79**	**93.08**	**91.57**	**77.00**	**96.74**	**92.13**	**88.01**	**91.26**	**90.70**	**82.85**	**97.06**	**90.64**
One-shot	SapBERT	69.56	87.22	76.34	63.84	93.29	78.93	75.00	90.68	81.21	73.51	94.13	83.59
	MTransE	0.0	0.16	0.0	0.05	35.09	0.16	0.0	0.28	0.14	0.27	28.89	0.37
	HiPrompt	**92.11**	**95.11**	**93.53**	**77.63**	**97.25**	**93.91**	**88.28**	**91.53**	**90.61**	**81.31**	**96.39**	**90.28**

Table 2:

Statistics of the KG-Hi-BKF benchmark.

Dataset	Source	#Disease	#Entities	#Links
SDKG-DzHi	SDKG	841	19,416	635
	DzHi	11,159	11,159	635
repoDB-DzHi	repoDB	2,074	3,646	709
	DzHi	11,159	11,159	709

Table 3:

Retriever with various expansion strategies.

Expan.	***SDKG-DzHi***			***repoDB-DzH***		
	Hits@5	**Hits@10**	**Hits@20**	**Hits@5**	**Hits@10**	**Hits@20**
Name	88.66	89.61	90.55	85.05	88.72	90.27
+Atr.	94.96	96.85	98.11	89.00	92.52	95.20
+Str.	90.08	90.71	91.81	88.15	90.27	92.24
+Atr.+Str.	**96.85**	**97.64**	**98.74**	**91.11**	**93.65**	**95.63**

Table 4:

Re-ranker with various LLMs and prompts.

LLMs	**SDKG-DzTaxo**			**repoDB-DzTaxo**		
	Hits@l	**Hits@3**	**MRR**	**Hits@l**	**Hits@3**	**MRR**
	One-shot (prompt w/o Hi. Context)					
GPT-3	**91.80**	**94.32**	**93.45**	**87.85**	**91.24**	**89.92**
GPT-JT	75.08	86.44	81.80	58.33	69.77	66.42
OPT-6.7B	68.93	80.44	76.38	60.73	73.59	69.33
	One-shot (prompt w/ Hi. Context)					
GPT-3	**92.11**	**95.11**	**93.91**	**88.28**	**91.53**	**90.28**
GPT-JT	80.76	93.69	87.45	69.07	82.91	77.24
OPT-6.7B	72.40	84.86	79.64	63.70	77.68	72.41