The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

Identifying Bridging Entity and Its Context using Dense Retriever

Junyoung Son°†, Jinsung Kim†, Jungwoo Lim†, Yoonna Jang†, Aram So*‡
Department of Computer Science and Engineering, Korea University†, Human-inspired AI Research‡
{s0ny,jin62304,wjddn803,morelychee,aram}@korea.ac.kr

Abstract

Cross-document relation extraction (CodRED) is a foundational task to build the knowledge base for enhancing many downstream tasks, such as question-answering and dialogue systems. Recently, CodRED, introduced for reasoning the density of knowledge distributed across multiple documents in the real world, has garnered significant attention. In this setup, the target entity pair, consisting of *head* and *tail*, is not co-mentioned within a single document. Thus, it needs to infer their relationship, considering the direct and indirect context between the documents, through intermediate entities that potentially connect the target entities. However, extracting such information from raw text is a non-trivial challenge. In this study, we focus on identifying an arbitrary entity and extracting its contexts that are potentially related to the target entities. We hope that this retriever can be employed in constructing bridging contexts further to improve the quality of reasoning paths in the CodRED.

Keywords: Cross-document Relation Extraction, Entity Linking, Information Retrieval

1. Introduction

Cross-document relation extraction (CodRED) is a foundational task to build the knowledge base for many Natural Language Processing (NLP) tasks, such as question-answering and interactive dialogue systems. Drawing upon the recent developments, CodRED, designed to simulate the sparsity of real-world knowledge dispersed across various documents, has garnered considerable interest and is the subject of ongoing research [1]. In this dataset, it is required to infer a set of reasoning paths that each of them consists of the pair of documents (d^h, d^t), where each document includes mentions of *head* or *tail* entity. To figure out shreds of related evidence to be helpful for inferring the relation between *head* and *tail* using the set of reasoning paths, it is important to detect bridging entities that might be connected *head* and *tail* entities with potential relational clues.

However, previous studies have only focused on detecting superficial reasoning paths by aggregating the number of (*head*, *tail*) mentions [1, 2] rather than detecting which entity is more important. To alleviate this problem, we propose an Entity-Centric Retriever (ECR) to find related bridging contexts to the target entities, which regard entity names as queries to retrieve related contexts. In our experiments, we observed 94.0% of Hits@1 in the test set, implying ECR's prominent capability to retrieve bridging contexts. We hope researchers can identify important reasoning paths by retrieving bridging entity-related contexts through ECR.

2. Method

2.1 Dataset

To the best of our knowledge, there is no dataset provided for training the retriever employing entity name as a query and their corresponding page as a positive document because most of the retrievers use a query that consists of semantic elements [3–5] to support a question answering task mainly. Thus, we need to construct a new dataset to train ECR. To this end, we utilize Wikipedia and Wikidata dumps to get entity-specific information and their corresponding document. In detail, we only consider the first passage of each document as our retrieval unit. In total, we collected 104,352 (*entity*, *contexts*) pairs for training. Since we aim to identify

*Corresponding author

Method	Development Set			Test Set		
	Hits@1	Recall@5	MRR	Hits@1	Recall@5	MRR
DPR	88.2	97.1	92.1	87.7	96.9	91.8
+ ID filtering	89.3 (+1.1)	97.4 (+0.3)	92.9 (+0.8)	88.9 (+1.2)	97.3 (+0.4)	92.7 (+0.9)
ECR	93.7	98.8	96.0	93.4	98.7	95.8
+ ID filtering	**94.4** (+0.7)	**99.0** (+0.2)	**97.0** (+1.3)	**94.0** (+0.6)	**98.7** (+0.0)	**96.2** (+0.4)

Table 1. Retrieval performances of ECR on the development and test sets.

the representative entity and its evidential contexts to connect *head* and *tail*, we filter out the entities with no connected Wikipedia page to ensure a one-to-one mapping.

2.2 Training

We employ a dense passage retriever (DPR) proposed by [6]. It consists of two encoder models: the first one is to encode the query, and the other is used to encode the contexts. To train ECR, we utilize the training objective following contrastive learning [7], as shown in Equation 1.

$$l(e,p^+,\{p_j\}_{j=1}^J) = -\log \frac{\exp^{sim(e,p_i^+)}}{\exp^{sim(e,p_i^+)} + \sum_{j=1}^J \exp^{sim(e,p_j^-)}}, \quad (1)$$

where p^+ is a positive document of entity e and $\{p_j^-\}_{j=1}^n$ is a set of negative documents that are extracted by BM25.

2.3 Bridging Entity Identification

To prevent the ECR from returning irrelevant contexts with the input entity, we apply a filtering strategy by using a unique entity ID based on Wikidata. In detail, we filter out the documents that do not contain the query entity ID from the set of candidate documents to ensure entity-related bridging contexts. This can be seen as entity filtering based on its retrieved contexts.

3. Experiment

3.1 Experimental setting

The statistics of the dataset utilized to train ECR are shown in Table 2.

To evaluate ECR's capabilities to retrieve entity-related documents, we utilize Hits@1/Recall@5/Mean Reciprocal Rank (MRR) as our evaluation metric, where Hits@1 is to measure the accuracy using the predicted top 1 document.

	Train	Dev	Test
Number of query-doc pairs	104,352	4,601	4,576

Table 2. Statistics of the dataset utilized to train ECR.

To compare our ECR to existing retrievers, we employ the DPR [6] model trained in Natural Questions (NQ) [5].

We collect all identified bridging entities from CodRED for each development and test set to verify ECR's capability to find bridging contexts in cross-document scenarios. In detail, we collect all bridging entities and align them by using Wikidata and Wikipedia.

3.2 Experimental result

As shown in Table 1, ECR shows the best performances across all metrics compared to DPR. In addition, we observed improvements with the ID filtering module, implying that disambiguating entities is essential to identifying bridging contexts.

4. Conclusion

In this paper, we proposed an Entity-Centric Retriever to provide bridging contexts between *head* and *tail* entities. With only the entity's name, we observed that ECR can effectively provide entity-related contexts. In our experiments, we verified ECR's applicability in CodRED dataset to find bridging contexts. We hope that this finding can offer additional perspectives to search relevant reasoning paths. We also note that to use ECR in constructing paths, a metric estimating the importance of the bridging entity may be required.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

Reference

[1] Y. Yao, J. Du, Y. Lin, P. Li, Z. Liu, J. Zhou, and M. Sun, "Codred: A cross-document relation extraction dataset for acquiring knowledge in the wild," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4452–4472, 2021.

[2] K. Lu, I. Hsu, W. Zhou, M. D. Ma, M. Chen *et al.*, "Multi-hop evidence retrieval for cross-document relation extraction," *arXiv preprint arXiv:2212.10786*, 2022.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Nov. 2016. [Online]. Available: https://aclanthology.org/D16-1264

[4] M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer, "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Jul. 2017. [Online]. Available: https://aclanthology.org/P17-1147

[5] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, "Natural questions: A benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 452–466, 2019. [Online]. Available: https://aclanthology.org/Q19-1026

[6] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Nov. 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.550

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *International conference on machine learning*, pp. 1597–1607, 2020.