Advancing Cross-Document Relation Extraction with Hybrid Retrieval and Knowledge-Augmented Reasoning

Marco Martinelli[1,*]

[1]Department of Information Engineering, University of Padova, Padova, Italy

Abstract

Existing Relation Extraction (RE) methods typically focus on extracting relational facts between entity pairs within single sentences or documents. However, in practice, a large amount of relational facts can only be inferred by reasoning across multiple documents. In this work, we introduce the task of Cross-Document Relation Extraction (CDRE), placed in between the domains of Information Retrieval (IR) and Natural Language Processing (NLP). CDRE enables the acquisition of knowledge in the wild, making it better suited for real-world use cases where relevant information is scattered across multiple sources. After formally introducing the task and the components involved in a CDRE system, we present the research directions that we plan to pursue to advance the state of the art. Specifically, we propose to integrate sparse and dense retrieval models with the heuristic-based methods currently employed in CDRE to improve the retrieval effectiveness of relevant passages from multiple documents. To further improve this retrieval, we introduce path-ranking algorithms as re-rankers to filter out less informative passages. Additionally, we explore leveraging graph-based representations to enhance document retrieval. Next, we plan to adapt Knowledge Injection (KI) techniques widely employed in sentence- and document-level RE to the CDRE setting, aiming to improve their robustness against syntactic and semantic variations, hence enhancing extraction effectiveness. Finally, we present an evaluation framework designed to assess the overall performances of CDRE systems and analyze the impact of each individual component.

Keywords

Natural Language Processing, Relation Extraction, Cross-Document Relation Extraction, Information Retrieval, Document Retrieval, Passage Retrieval

1. Introduction

RE is a crucial NLP task that aims to detect the semantic relations between a pair or target entities in a given text. It is fundamental for natural language understanding, automated Knowledge Bases (KBs) construction and population, and, more generally, for nearly all knowledge-driven AI tasks [1, 2].

Most of the RE research has been limited to scenarios where the entity pair is within a single sentence or document [2]. However, many use cases require extracting relations across multiple documents, where the target entity pair may not coexist within the same document. For example, more than 57.6% of the relational facts in Wikidata are not described in individual Wikipedia documents [3].

That led to the introduction of the CDRE task, which breaks through the limitations of document boundaries to acquire knowledge from multiple text sources. CDRE requires a RE system to infer the relation holding between two entities by retrieving and reasoning over multiple documents from a large-scale corpus of documents. Therefore, CDRE places itself in the middle of the IR and NLP domains.

Compared to more traditional RE tasks, CDRE introduces two main challenges. The first relates to Document Retrieval (DR) and Passage Retrieval (PR), requiring systems to identify relevant documents and extract from them informative passages for the target entity pair. The second difficulty is associated with reasoning, since working at the cross-document level requires systems to perform both intra- and cross-document reasoning across multiple documents and then predict the relations by aggregating information [3].

SEBD 2025: 33rd Symposium on Advanced Database System, June 16-19, 2025 - Ischia, Italy

[*]Corresponding author.

martinell2@dei.unipd.it (M. Martinelli)
https://www.dei.unipd.it/~martinell2/ (M. Martinelli)
0009-0001-1596-8642 (M. Martinelli)

In this research project, the primary effort will be on improving the retrieval component of CDRE systems, using dense and sparse retrievers typically employed in IR and path-mining and ranking approaches used in Open-Domain Question Answering (ODQA). Although these techniques have shown great potential in their specific domains, their synergy with CDRE is still in its early stages. Improvements can be made in matching relevant passages scattered across multiple documents, balancing syntactic and semantic aspects to ensure that relevant information is captured both for individual entities and for the entity pair as a whole, and in optimizing the ranking specifically for the CDRE task, prioritizing the most informative passages for relation inference.

Furthermore, we will model DR in CDRE using knowledge graphs to implement a hybrid approach that integrates traditional keyword-based methods with knowledge graph-based ones.

The next step will be to employ KI, which is incorporating external information (e.g., KBs and ontologies) into CDRE models to improve their reasoning and inference capabilities. This technique has been widely employed in both sentence- and document-level RE, showing great effectiveness [4, 5]. However, the integration of these techniques at the cross-document level is more difficult since managing context and ensuring that injected knowledge remains accurate and pertinent throughout different documents presents significant challenges.

The following sections are structured as follows: Section 2 describes the CDRE task in detail and presents the current methods and resources for CDRE; Section 3 details the proposed research questions and directions; and Section 4 concludes with some final remarks.

2. Task Description

Given a target entity pair (h, t) a (large-scale) corpus of documents D, and a set of relations $R = \{r_1, \ldots, r_n\} \cup$ N/A, the CDRE task can be decomposed into four stages:

1. **Document Retrieval (DR)**, which requires the system to find documents $\{d_1, \ldots, d_i\} \in D$ relevant to the pair (h, t).
2. **Passage Retrieval (PR)**, which extracts the most informative passages $\{p_1, \ldots, p_j\} \in \{d_1, \ldots, d_i\}$ to ensure that the input fits within the token limit of the employed RE model.
3. **Input Construction (IC)**, that is, processing the selected passages into a format suitable for RE. This usually involves combining the selected passages into a single text representation, hence reconducing the CDRE task to a document-level RE problem.
4. **Relation Extraction (RE)**, consisting of reasoning over the provided input to predict the relation $r \in R$ between h and t.

In that scope, we define an entity e as a *bridge entity* if it is in relation with h in a document d_x and with t in a document d_y. Furthermore, a pair of documents (d_x, d_y) containing, respectively, the head ($h \in d_x$) and tail ($t \in d_y$) entities and connected by a bridge entity is called *(reasoning) text path*.

In Figure 1, the documents "*Pink Floyd: Live at Pompeii*" and "*Progressive music*" represent a text path, with the bridging entities that connect the passages linked with arrows.

2.1. Related Works

Currently, there is only one dataset specifically designed to test the RE systems' ability at the cross-document level: CodRED [3]. It includes over 30'000 positive relations associated with more than 210'000 documents, providing ground truth for both DR and CDRE. Furthermore, CodRED features a subset of relations annotated with evidence sentences, namely sentences belonging to the head and tail documents that allow the inference of the annotated relation, providing ground truth for PR.

CodRED includes two benchmark settings to fully evaluate each required capability of a CDRE system:

1. **Closed setting**: The model is provided with the entity pair (h, t) and the corresponding text path (d_h, d_t) from which to establish the relation.

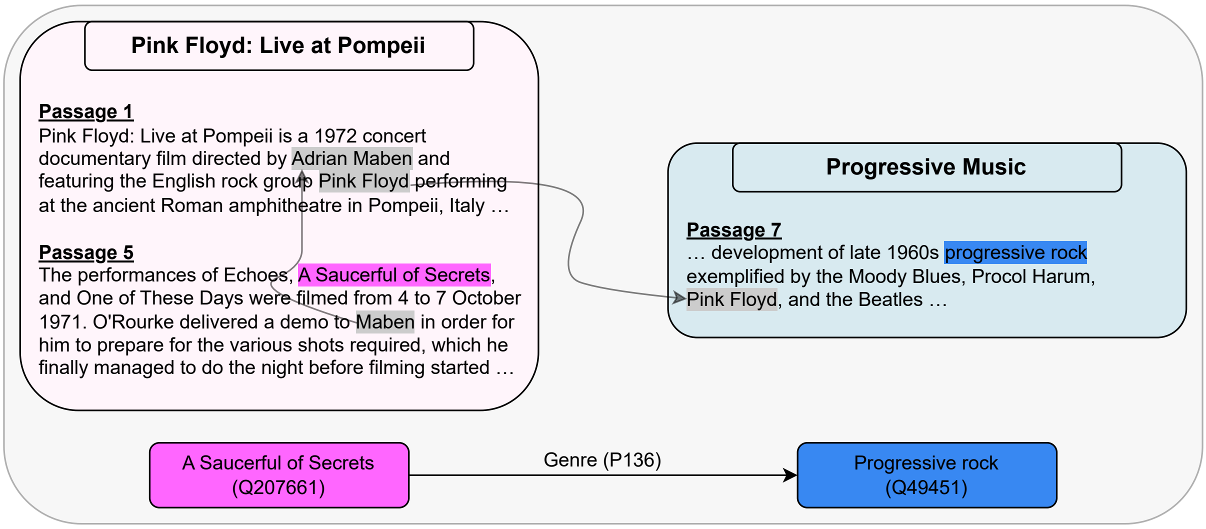
Pink Floyd: Live at Pompeii
Passage 1
Pink Floyd: Live at Pompeii is a 1972 concert documentary film directed by Adrian Maben and featuring the English rock group Pink Floyd performing at the ancient Roman amphitheatre in Pompeii, Italy …
Passage 5
The performances of Echoes, A Saucerful of Secrets, and One of These Days were filmed from 4 to 7 October 1971. O'Rourke delivered a demo to Maben in order for him to prepare for the various shots required, which he finally managed to do the night before filming started …
Progressive Music
Passage 7
… development of late 1960s progressive rock exemplified by the Moody Blues, Procol Harum, Pink Floyd, and the Beatles …
A Saucerful of Secrets (Q207661)
Genre (P136)
Progressive rock (Q49451)

Figure 1: Example of CDRE showing how the relation *(A Saucerful of Secrets, Genre, Progressive rock)* is extracted from three passages scattered across two Wikipedia documents.

2. **Open setting**: The model is provided only with the entity pair (h, t) and needs to first retrieve relevant documents from the corpus of documents before determining the relation.

Previous works have focused on one or both of these settings, implementing one or more of the stages required in CDRE, as summarized in Table 1.

Table 1
Overview of previous works in CDRE.

Model	Paper	Year	Documents Retrieval	Input Construction	Relation Extraction	Closed Setting	Open Setting
CodRED Baselines	[3]	2021	✓	✓	✓	✓	✓
ECRIM	[6]	2022	✗	✓	✓	✓	✗
MR.COD	[2]	2023	✓	✓	✗	✓	✓
PILOT	[7]	2023	✓	✗	✓	✓	✗
KXDocRE	[8]	2024	✗	✓	✓	✓	✓
REIC	[9]	2024	✗	✓	✗	✓	✗
N-B EE+PD	[10]	2024	✗	✓	✓	✓	✓

3. Research Directions

This project aims to integrate techniques from the IR, ODQA, and KI domains into CDRE, addressing two interlinked yet distinct objectives: enhance the DR, PR, and IC stages to improve extraction independently of the employed downstream RE model, and advance the cross-document reasoning capabilities of RE models to advance the state of the art (SOTA) in terms of effectiveness. To achieve these goals, three Research Questions (RQs) have been outlined:

RQ1: Are dense and sparse retrieval methods, along with path-ranking algorithms, effective for retrieving relevant passages for CDRE? Is it appropriate to apply knowledge graphs for DR in CDRE?

RQ2: Can KI enhance the robustness of CDRE models against syntactic and semantic variations and improve their flexibility to adapt to varying relation descriptions across different contexts?

RQ3: How to evaluate the effectiveness of the proposed CDRE systems and the individual impact of the employed methods on the overall CDRE effectiveness?

3.1. Research Design & Methods

Based on the defined RQs, the research framework illustrated in Figure 2 has been outlined.

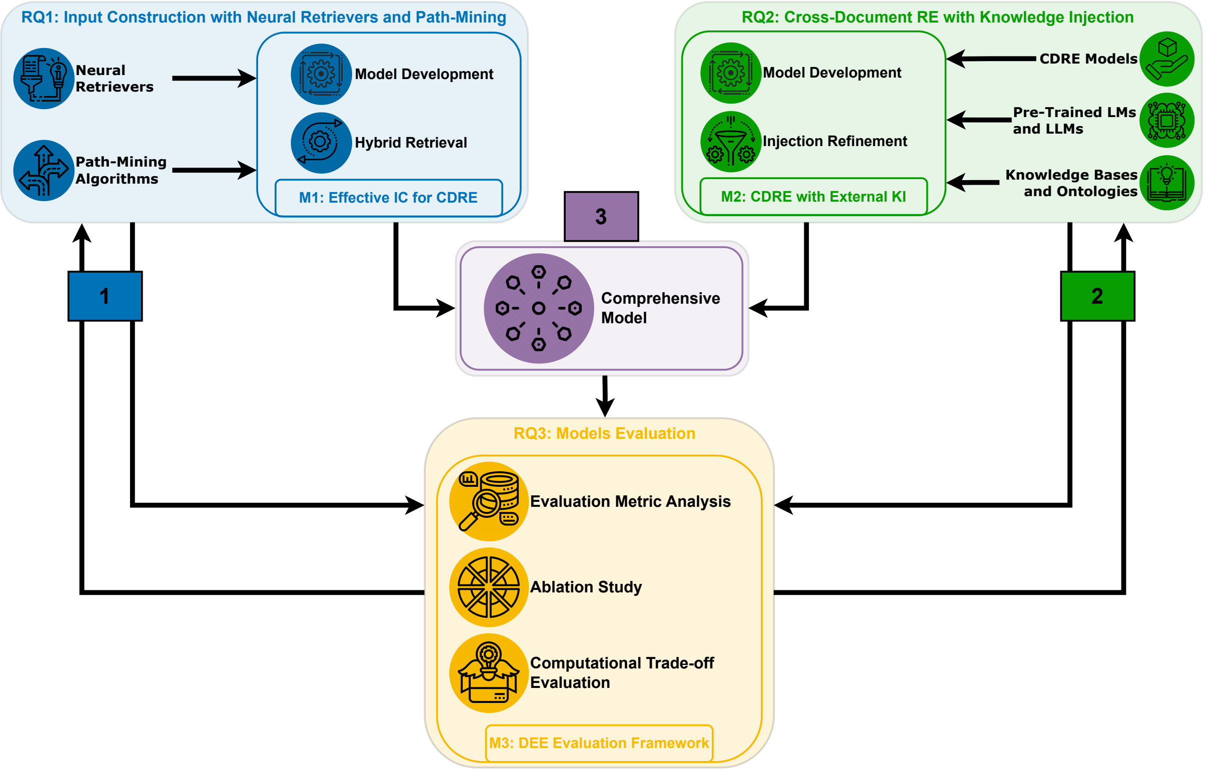
RQ1: Input Construction with Neural Retrievers and Path-Mining
Neural Retrievers
Path-Mining Algorithms
Model Development
Hybrid Retrieval
M1: Effective IC for CDRE
RQ2: Cross-Document RE with Knowledge Injection
Model Development
Injection Refinement
M2: CDRE with External KI
CDRE Models
Pre-Trained LMs and LLMs
Knowledge Bases and Ontologies
1
2
3
Comprehensive Model
RQ3: Models Evaluation
Evaluation Metric Analysis
Ablation Study
Computational Trade-off Evaluation
M3: DEE Evaluation Framework

Figure 2: CDRE research project framework based on the defined RQs.

Regarding **RQ1** (see the blue rectangle), the first step is to reproduce the heuristic-based methods proposed in [6] and [8] and integrate the scores from different PR models, such as SPLADE [11], Contriever [12], DPR [13], and ColBERT [14], into these heuristics. The inputs constructed in this manner will then be fed into the same downstream RE model to assess their different impacts on extraction effectiveness.

Next, we will select the PR models that demonstrate the best performance and replace the heuristic-based methods with path-ranking algorithms. Specifically, we will first retrieve relevant passages and then rank them using these algorithms, extending the approach proposed in [7] by constructing a knowledge graph of passages and scoring their importance based on their centrality within the graph [15].

Finally, we will implement a hybrid DR approach that integrates traditional keyword-based methods, such as BM25 and TF-IDF, with knowledge-graph based techniques. Given an entity pair (h, t), we will first construct a knowledge graph where documents are nodes and edges represent bridging entities, shared entities, and semantic similarity [2]. These edges will be weighted, with weights potentially learnable. We will then estimate the centrality of documents within this graph and use it as a relevance score [16, 17], which will be combined with the traditional scoring functions mentioned above to select the most relevant documents.

To answer **RQ2** (see the green rectangle), we will enhance the text understanding in CDRE models by augmenting texts with detailed attributes and relationships from domain-specific ontologies and KBs.

We will address imperfections and coverage gaps of KBs by extending techniques applied for automatic KBs enrichment, such as random walks [18] and rule mining models [19], to infer missing details in

KBs and enhance the accuracy of augmented texts.

Subsequently, we will address the challenge of adapting to relation descriptions and syntactic and semantic variations by adapting techniques from RE tasks that employ pre-trained Language Models (LMs) [20] and Large Language Models (LLMs) [21], known for their effectiveness in capturing long-distance relations and mapping entities to unique identifiers [22]. Furthermore, by incorporating KI, we aim to enhance the robustness and context awareness of these models, reducing their susceptibility to hallucinations and insensitivity to negations.

To address **RQ3** (see yellow rectangle), we will first evaluate the impact of the various PR and IC approaches on the effectiveness of the downstream CDRE models.

Following the identification of the most effective ones, we will integrate them into a single model (i.e., RQ1) and conduct an ablation study to assess their individual contributions. These approaches will then be applied to construct the inputs for CDRE models and LMs with external KI (i.e., RQ2), with another round of iterative evaluation to refine their application.

We will then evaluate the effectiveness of the developed DR methods by comparing the RE scores obtained for the same entity pairs in both the open and closed settings, allowing us to determine whether the retrieved documents enable the models to predict the same relations they identify when provided with the exact document pair and, consequently, understand the impact of DR on the overall performance of CDRE systems.

For evaluation, DR and PR performances will be measured using standard IR evaluation metrics: Precision@K, Recall@K, F1-score, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG). Concerning RE, since it can be considered a multi-label classification problem, Precision, Recall, and F1-score will be used for evaluation.

4. Conclusion

The CDRE task lies in between the IR and NLP domains, presenting a wide range of challenges at different levels of granularity. Despite the modernity of the task, various promising approaches have been proposed in the literature, implementing one or more of the elements involved in a CDRE system.

However, there is significant room for improvement in all components of CDRE, along with a need to better understand their individual impact on overall effectiveness. Therefore, the primary foundational step of this project involves working on the retrieval and IC layers by reproducing and augmenting the solutions currently developed and evaluating their impact on a fixed downstream RE model. After that, the focus of the project will be on improving the reasoning and inference capabilities of CDRE models by leveraging external KI.

The final goal of my research is the release of an end-to-end CDRE system that can work in both closed and open settings, improving the current SOTA and potentially enhancing the integration between Document Retrieval, Passage Retrieval, Input Construction, and reasoning to achieve more robust and effective Cross-Document Relation Extraction.

Acknowledgments

This project has received funding from the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

Declaration on Generative AI

During the preparation of this work, the author used GPT-4o and Grammarly in order to: Grammar and spelling check. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. Ali, M. Saleem, D. Moussallem, M. A. Sherif, A.-C. Ngonga Ngomo, Reld: A knowledge graph of relation extraction datasets, in: C. Pesquita, E. Jimenez-Ruiz, J. McCusker, D. Faria, M. Dragoni, A. Dimou, R. Troncy, S. Hertling (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2023, pp. 337–353.

[2] K. Lu, I.-H. Hsu, W. Zhou, M. D. Ma, M. Chen, Multi-hop evidence retrieval for cross-document relation extraction, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, p. 10336–10351. URL: https://aclanthology.org/2023.findings-acl.657. doi:10.18653/v1/2023.findings-acl.657.

[3] Y. Yao, J. Du, Y. Lin, P. Li, Z. Liu, J. Zhou, M. Sun, Codred: A cross-document relation extraction dataset for acquiring knowledge in the wild, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, p. 4452–4472. URL: https://aclanthology.org/2021.emnlp-main.366. doi:10.18653/v1/2021.emnlp-main.366.

[4] X. Wang, Z. Wang, W. Sun, W. Hu, Enhancing document-level relation extraction by entity knowledge injection, in: The Semantic Web – ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2022, p. 39–56. URL: https://doi.org/10.1007/978-3-031-19433-7_3. doi:10.1007/978-3-031-19433-7_3.

[5] B. Chen, G. Yuan, Domain knowledge-driven relation extraction methods, Journal of Data Science and Intelligent Systems (2024). URL: https://ojs.bonviewpress.com/index.php/jdsis/article/view/2524. doi:10.47852/bonviewJDSIS42022524.

[6] F. Wang, F. Li, H. Fei, J. Li, S. Wu, F. Su, W. Shi, D. Ji, B. Cai, Entity-centered cross-document relation extraction, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, p. 9871–9881. URL: https://aclanthology.org/2022.emnlp-main.671. doi:10.18653/v1/2022.emnlp-main.671.

[7] J. Son, J. Kim, J. Lim, Y. Jang, H. Lim, Explore the way: Exploring reasoning path by bridging entities for effective cross-document relation extraction, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, p. 6755–6761. URL: https://aclanthology.org/2023.findings-emnlp.450. doi:10.18653/v1/2023.findings-emnlp.450.

[8] M. Jain, R. Mutharaju, K. Singh, R. Kavuluru, Knowledge-driven cross-document relation extraction, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 3787–3797. URL: https://aclanthology.org/2024.findings-acl.227/. doi:10.18653/v1/2024.findings-acl.227.

[9] B. Na, S. Jo, Y. Kim, I.-c. Moon, Reward-based input construction for cross-document relation extraction, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, p. 9254–9270. URL: https://aclanthology.org/2024.acl-long.501. doi:10.18653/v1/2024.acl-long.501.

[10] H. Yue, S. Lai, C. Yang, L. Zhang, J. Yao, J. Su, Towards better graph-based cross-document relation extraction via non-bridge entity enhancement and prediction debiasing, in: Findings of the Association for Computational Linguistics ACL 2024, Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2024, p. 680–691. URL: https://aclanthology.org/2024.findings-acl.38. doi:10.18653/v1/2024.findings-acl.38.

[11] T. Formal, B. Piwowarski, S. Clinchant, Splade: Sparse lexical and expansion model for first stage ranking, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Virtual Event Canada, 2021, p. 2288–2292. URL: https://dl.acm.org/doi/10.1145/3404835.3463098. doi:10.1145/3404835.3463098.

[12] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense

information retrieval with contrastive learning, CoRR (2021). URL: https://arxiv.org/abs/2112.09118. doi:10.48550/ARXIV.2112.09118.

[13] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, p. 6769–6781. URL: https://www.aclweb.org/anthology/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.

[14] K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia, Colbertv2: Effective and efficient retrieval via lightweight late interaction, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, p. 3715–3734. URL: https://aclanthology.org/2022.naacl-main.272. doi:10.18653/v1/2022.naacl-main.272.

[15] R. Ribeiro, D. M. de Matos, Revisiting centrality-as-relevance: support sets and similarity as geometric proximity, J. Artif. Int. Res. 42 (2011) 275–308.

[16] W. Kim, H. Jang, H.-J. Kim, D. Kim, A document query search using an extended centrality with the word2vec, in: Proceedings of the 18th Annual International Conference on Electronic Commerce: E-Commerce in Smart Connected World, ICEC '16, Association for Computing Machinery, New York, NY, USA, 2016. URL: https://doi.org/10.1145/2971603.2971617. doi:10.1145/2971603.2971617.

[17] O. Levi, F. Raiber, O. Kurland, I. Guy, Selective cluster-based document retrieval, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1473–1482. URL: https://doi.org/10.1145/2983323.2983737. doi:10.1145/2983323.2983737.

[18] N. Lao, T. Mitchell, W. W. Cohen, Random walk inference and learning in a large scale knowledge base, in: R. Barzilay, M. Johnson (Eds.), Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011, pp. 529–539. URL: https://aclanthology.org/D11-1049/.

[19] L. A. Galárraga, C. Teflioudi, K. Hose, F. Suchanek, Amie: association rule mining under incomplete evidence in ontological knowledge bases, in: Proceedings of the 22nd international conference on World Wide Web, ACM, Rio de Janeiro Brazil, 2013, p. 413–422. URL: https://dl.acm.org/doi/10.1145/2488388.2488425. doi:10.1145/2488388.2488425.

[20] A. Romadhony, D. H. Widyantoro, A. Purwarianti, Utilizing structured knowledge bases in open ie based event template extraction, Applied Intelligence 49 (2019) 206–219. doi:10.1007/s10489-018-1269-0.

[21] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, E. Chen, Large language models for generative information extraction: A survey, arXiv preprint arXiv:2312.17617 (2023).

[22] J. H. Caufield, H. Hegde, V. Emonet, N. L. Harris, M. P. Joachimiak, N. Matentzoglu, H. Kim, S. Moxon, J. T. Reese, M. A. Haendel, P. N. Robinson, C. J. Mungall, Structured prompt interrogation and recursive extraction of semantics (spires): a method for populating knowledge bases using zero-shot learning, Bioinformatics 40 (2024) btae104. doi:10.1093/bioinformatics/btae104.