중간 문맥 식별 및 검색을 활용한 문서간 관계 추출

손준영°†, 김진성†, 임정우†, 장윤나†, 소아람*§, 임희석*†§
고려대학교 컴퓨터학과†, Human-inspired AI 연구소§
{s0ny,jin62304,wjddn803,morelychee,aram,limhseok}@korea.ac.kr

Cross-document Relation Extraction using Bridging Context Identification

Junyoung Son°†, Jinsung Kim†, Jungwoo Lim†, Yoonna Jang†, Aram So*§, Heuiseok Lim*†§
Department of Computer Science and Engineering, Korea University†, Human-inspired AI Research§

요 약

관계 추출은 질의응답이나 대화 시스템의 기반이 되는 지식을 구추하기 위한 작업으로, 정보 추출의 기초가 되는 기술이기도 하다. 최근 실세계 지식의 희소한 형태를 구현한 문서간 관계 추출 데이터셋이 제안되어, 여러 문서를 통해 분산되어 언급된 두 개체 사이의 관계 추론을 수행 및 평가할 수 있게 되었다. 이 작업에서 추론의 대상이 되는 개체쌍은 한 문서 안에 동시에 언급되지 않기 때문에 여러 문서에 언급된 중간 개체를 통하여 직/간접적으로 추론해야 하나, 원시 텍스트에서 이러한 정보를 수집하는 작업은 쉽지 않다. 따라서, 본 연구에서는 개체의 동시발생빈도에 기반하여 중간 개체의 중요도를 정량화하고, 이 중요도에 기반화여 중요한 문맥을 식별하는 방법론을 제안한다. 제안하는 방법론은 기존의 두 문서로 구성된 추론 경로를 식별된 중간 개체를 활용하여 확장하여, 관계 추론 모델의 수정 없이 추가된 문맥만을 활용하여 문서간 관계 추출 시스템의 성능을 개선할 수 있었다.

주제어: 관계 추출, 지식 추출, 문서 관계 추출, 문서간 관계 추출

1. 서론

최근 Large Language Models (LLMs)가 큰 이슈가 되면서, 이들의 추론 능력에 검색이나 지식을 융합하려는 다양한 시도가 있어 왔다 [1, 2]. 즉, 질의응답이나 대화시스템과 같이 지식베이스에 의존하는 기술은 예전과 다르지 않게 아직까지 현존하며, 이러한 기대에 부응하기 위하여 지식베이스를 구축하고 확장하는 기술은 필수적이라고 볼 수 있다 [3, 4]. 특히, 관계 추출은 임의의 두 개체 사이의 관계를 추론하여 *(주체, 관계, 개체)*와 같이 구조화된 지식을 구축할 수 있기 때문에, 개체명 인식이나 개체 연결 작업과 함께 지식 추출 작업의 근간이 되는 기술로 여겨져 왔다 [5].

기록된 여러 문서에서 한 문서에 같이 언급되지 않는 임의의 두 개체 사이의 관계를 추론하는 상황을 가정해보자. 두 개체가 한 문서 내에 같이 언급되지 않더라도 서로 다른 여러 문서를 통해서 암시적으로 관계를 유추할 수 있을 것이다. 앞선 예측과는 다르게, 관계 추출 상황을 재현한 여러 데이터셋들은 한정된 상황만을 가정한다. 즉, 이들은 단일 문장이나 단일 문서 혹은 단일 대화와 같이 좁은 범위의 지식만을 고려하도록 제안되었다. 그러나, 방대한 지식베이스인 위키데이터에 저장된 지식의 통계를 보면 약 57.6%의 지식이 단일 문서 안에서 언급되지 않는다 [6]. 대표적인 지식베이스에 저장된 절반 이상의 지식이 여러 문서에 걸쳐서 분산되어 있다는 사실은 기존의 관계 추출이 실제 세상의 지식을 모두 포괄하지 못할 수 있음을 의미한다. 즉, 이러한 형태의 지식을 고려한 문서간 관계 추출에 대한 연구가 필요한 시점이다.

문서간 관계 추출은 여러 하위 작업으로 구성된 복잡한 추론을 요구하는 작업이다. 첫 번째 작업은 주어진 개체 쌍과 관련된 문서를 수집하는 것이다. 이 작업은 주체 개체와 객체 개체의 각각의 관점에서 두 개체의 관계를 추론하기 위한 문서를 수집하는 작업이다. 각 개체 관점의 문서 집합은 이후 임의의 중간 개체로 연결된 추론 경로를 구축하기 위해 사용된다. 다음 단계 작업은 구축된 추론 경로 집합에서 두 개체의 관계 추론에 필요한 텍스트를 식별 및 수집하는 것이다. 문서는 일반적으로 많은 단락으로 구성되기 때문에 일반적인 언어 모델(BERT [7], RoBERTa [8])의 입력 범위를 벗어나기 때문에, 한정된 입력 안에 밀집된 정보를 제공하여 관계 추론을 극대화할 필요가 있다.

이러한 어려움을 해소하기 위해서 본 연구는 동시출현빈도에 기반하여 추론 경로 내의 개체의 중요도를 정량화하고, 이 정량화된 점수에 기반하여 개체와 관련있는 추가적인 문맥을 검색하는 연구를 수행한다. 즉, 두 문서로 구성된 추론 경로 내에 등장하는 중간 개체의 점수를 정량화하고 이들을 활용하여 중간 문맥을 검색하고, 이 문맥을 활용하여 문서간 관계 추출을 수행한다. 실험 및 분석을 통해서 식별된 문맥이 추가된 추론 경로가 기존의 추론 경로를 활용한 모델보다 좋은 성능을 보여 제안하는 방법론의 효과성을 입증하였다.

*교신저자(Corresponding author)

2. 관련 연구

관계 추출 작업은 정보 추출 및 지식 베이스 구축의 기반이 되는 기술로 여겨져 왔다. 문장 수준 관계 추출은 단일 문장 내에 언급된 두 개체 사이의 관계를 식별하는 작업으로, 다양한 연구가 활발하게 이루어져 왔다 [9]. [10]은 합성곱 신경망을 활용하여 문장과 어휘 수준 자질을 추출하는 기술을 적용하여 문장 수준 관계 추출을 수행하였다. [11]는 두 개체 사이의 관계의 방향성을 분류하는 새로운 방향의 연구를 제안한 바 있다. [12]와 같이 그래프 합성곱 신경망을 활용하여 문장 내에서 중요한 정보를 식별하려는 시도가 있었다.

최근 몇 년 동안 문서 단위 관계 추출은 많은 연구자들의 관심을 받아왔다 [13,14]. 특히, 단일 문서 관계 추출은 한 문서 안에 등장하는 개체 집합에 대하여 관계를 추출하는 작업이다. [15]는 문서와 문장, 그리고 개체 등 여러 유형을 고려한 정점과 간선을 활용한 문서 수준 그래프를 구축하여 연구를 수행한 바 있으며, []는 로컬 및 글로벌 추론에 기반한 맨션 기반 추론 모듈을 제안하였다.

CodRED라 불리는 문서간 관계 추출 데이터셋이 공개된 후, 다양한 관점의 연구가 진행되어 왔다. CodRED는 두 문서로 구성된 추론 경로의 집합을 활용하여 두 개체 사이의 관계 추론을 요구한다. [16]는 기존의 모델이 단일 추론 경로 내의 문서간 상호 작용만을 고려하는 점을 지적하여, 추론 경로 사이의 상호 작용을 모델링하여 연구를 진행한 바 있다. [17]는 긴 문서에서 정보를 선택하는 작업의 어려움을 완화하기 위하여 문서를 문단 수준으로 전처리하고, 중간 개체에 기반한 그래프 순회를 적용하여 문단 단위 추론 경로를 구축한 뒤, 재순위화 모델을 활용하여 최종 문단을 추출하였다.

3. 방법론

본 논문에서는 중간 개체의 주체 및 객체인 개체와의 동시출현빈도에 기반한 중요도를 정량화한 이전 연구 [16]에서 영감을 받아, 문서 수준 관계 추출을 위하여 추론 경로 내 문서에서 중요한 중간 개체를 찾고, 이를 기반으로 두 개체의 관계 추론에 유용한 문맥을 찾는데 집중한다. 제안하는 방법론은 추론 경로 내에서 중요한 개체를 식별하기 위한 모듈과, 각 개체와 관련된 문맥을 검색하는 검색 모델로 구분된다. 관계 추론 모델 [6]은 기존의 모델을 활용하였다.

3.1 중간 문맥 검색 모델

개체가 입력으로 주어졌을 때 관련된 문맥을 검색하기 위해서 본 연구에서는 사전학습된 밀집 검색 모델 [18]를 활용한다. 검색 대상으로는 CodRED 데이터셋에 주어지는 위키피디아 덤프 파일을 Faiss library [19]를 활용하여 Maximum Inner Product Search를 수행한다. 즉, 개체가 쿼리로 주어지면 전체 위키피디아 문서 집합에서 관련있는 문맥을 반환한다.

3.2 중간 개체 식별

주체($head$)와 객체($tail$)에 대한 두 문서로 구성된 추론 경로 $p = (doc_{head}, doc_{tail})$와 대응하는 추론 경로 내 개체 집합 (E_{head}, E_{tail})이 주어지면, 두 문서 개체 집합의 교집합에 대하여 중간 개체 후보군으로 사용한다. 개체 후보군에서 보다 유의미한 중간 개체를 식별하기 위해, 앞서 설명한 중간 문맥 검색 모델을 활용한다. 중간 문맥 검색 모델은 신경망 기반의 밀집 검색 모델이기 때문에, 개체의 포함 여부는 확정되지 않는다. 즉, 검색된 문맥이 쿼리로 활용된 중간 개체를 포함하지 않을 가능성이 있다. 우리는 이러한 현상을 보이는 후보군 개체가 유의미하지 않다고 판단하여 후보군 목록에서 제외하였다.

3.3 개체 중요도 정량화

단일 추론 경로 $p = (doc_{head}, doc_{tail})$에 존재하는 임의의 개체 e와 검색된 문맥 d_e을 활용하여 개체의 중요도는 아래와 같이 계산된다.

$$s(e) = (N(e, doc_{head}) + N(head, doc_e)) * (N(e, doc_{tail}) + N(tail, doc_e)), \quad (1)$$

$N(e, doc_i)$는 문서 doc_i에 개체 e가 언급된 횟수이다. 즉, 임의의 개체 e의 중요도를 검색된 문맥과 기존의 추론 경로에 기반한 동시출현빈도를 활용하여 평가한다.

4. 실험

4.1 데이터셋

실험에 사용한 데이터셋은 대표적인 문서간 관계 추출 데이터셋인 CodRED [6]이다. 데이터셋의 통계는 다음과 같다.

	Train	Dev	Test
# Positive facts	2,733	1,010	1,012
# N/A facts	16,668	4,558	4,523
# Reasoning paths	129,548	40,740	40,524

표 1. CodRED 데이터셋 통계. # Positive facts는 실제 관계가 존재하는 개체 쌍의 수를 의미하고, # N/A facts는 실제 관계가 존재하지 않는 개체 쌍의 수를 의미한다. # Reasoning paths는 개체 쌍에 대한 관계 추론에 사용되는 문서 집합의 수를 의미한다.

4.2 실험 환경

관계 추론 모델은 기존의 모델 [6]를 활용하기 때문에, 모든 하이퍼파라미터를 동일하게 활용하였다. 단, 학습 및 평가에

Method	Development set				Test set	
	Accuracy	F1 score	Precision@500	Precision@1000	Accuracy	F1 score
End-to-End [6]	47.94	51.26	62.80	51.00	47.46	51.02
+ Our method	51.87	55.74	72.26	55.44	53.97	57.54

표 2. 제안하는 방법론의 적용 유무에 따른 문서간 관계 추출 모델의 성능

활용한 추론 경로 데이터는 제안하는 방법론으로 확장한 것을 활용하여 진행하였다. 모델 학습에는 NVIDIA A6000 48GB 4개를 사용하였다. 관계 추론 모델을 위한 기반 언어 모델은 Huggingface [20]에서 제공하는 BERT-base-cased 모델을 활용하였다.

4.3 평가 기준

실험 성능 평가는 데이터셋 논문에서 제공하는 기준과 동일한 것(Accuracy, F1 score, Precision@500, Precision@1000)을 활용하였다 [6]. 테스트 세트의 경우 공식 리더보드에 제출해야 성능을 확인할 수 있기 때문에, Codalab에서 제공되는 CodRED의 리더보드를 통해서 평가하였다.

4.4 실험 결과

실험 결과는 표 2과 같다. 기존의 추론 경로로 학습 및 평가한 모델과 비교하여 큰 성능 개선을 보여, 본 논문에서 제안하는 방법론의 효과성을 입증하였다. 놀라운 것은 제안하는 방법을 적용한 모델은 테스트 세트에서 오히려 더 높은 성능을 보인다는 사실이다. 이 결과는 제안하는 중간 개체 식별 및 문맥 검색을 통한 추론 경로 확장 방법론의 높은 일반화 능력을 입증하는 것이라고 볼 수 있다. 또한, 이 결과를 다른 관점으로 해석할 수 있다면, 우리는 문서간 관계 추출에서 추론 경로라고 정의된 두 문서만 활용하는 것이 아니라, 두 개체 사이에 중간 다리역할을 하는 정보를 찾고 활용하는 것이 중요한 것이라는 점이다.

5. 결론

본 논문에서는 문서간 관계 추출을 개선화기 위한 중간 개체 식별 및 정량화, 문맥 검색 기술을 활용하였다. 제안하는 방법론은 관계 추론 모델의 수정 없이 입력으로 활용되는 추론 경로에 관련된 문맥을 추가함으로써 기존의 추론 경로를 효과적으로 개선하였다. 특히 테스트 세트에서의 높은 성능은 제안하는 방법론의 높은 일반화 성능을 말해주고 있다. 실험을 통해 알 수 있었던 점은 문서간 관계 추출에서 주어진 추론 경로만을 활용하는 것이 아니라, 중요한 정보를 어떻게 식별하고 어디에 활용할 지를 고안하는 것이 중요하다는 것이다. 본 연구의 한계점으로는 개체 중요도를 정량화할 때 사용하는 방식이 단순히 추론 경로 내 언급 횟수로 계산된다는 점인데, 직접 언급되지는 않지만 암시적으로 유용한 정보를 고려하지 못할 가능성이 있으며, 추론 경로조차 주어지지 않는 오픈 월드 관계 추출에서는 적용하기에 한계가 있다는 점에서 추론 경로의 양과 질에 의존적이라고도 볼 수 있다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2022-2018-0-01405). 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425). 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-00369, (4세부) 전문지식 대상 판단결과의 이유/근거를 설명가능한 전문가 의사결정 지원 인공지능 기술개발).

참고문헌

[1] R. Zhao, X. Li, S. Joty, C. Qin, and L. Bing, "Verify-and-edit: A knowledge-enhanced chain-of-thought framework," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5823–5840, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.320

[2] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," *The Eleventh International Conference on Learning Representations*, 2022.

[3] K. Swampillai and M. Stevenson, "Inter-sentential relations in information extraction corpora," *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

[4] H. Wu, X. Chen, Z. Hu, J. Shi, S. Xu, and B. Xu, "Local-to-global causal reasoning for cross-document relation extraction," *IEEE/CAA Journal of Automatica Sinica*, Vol. 10, No. 7, pp. 1608–1621, 2023.

[5] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, and M. Sun, "Docred: A large-scale document-level relation extraction dataset," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 764–777, 2019.

[6] 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.

[7] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681, 1997.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[9] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the tac 2010 knowledge base population track," *Third text analysis conference (TAC 2010)*, Vol. 3, No. 2, pp. 3–3, 2010.

[10] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2335–2344, Aug. 2014. [Online]. Available: https://aclanthology.org/C14-1220

[11] R. Cai, X. Zhang, and H. Wang, "Bidirectional recurrent convolutional neural network for relation classification," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 756–765, Aug. 2016. [Online]. Available: https://aclanthology.org/P16-1072

[12] Y. Zhang, P. Qi, and C. D. Manning, "Graph convolution over pruned dependency trees improves relation extraction," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2205–2215, Oct.-Nov. 2018. [Online]. Available: https://aclanthology.org/D18-1244

[13] H. Yang, D. Sui, Y. Chen, K. Liu, J. Zhao, and T. Wang, "Document-level event extraction via parallel prediction networks," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6298–6308, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.acl-long.492

[14] N. Zhang, X. Chen, X. Xie, S. Deng, C. Tan, M. Chen, F. Huang, L. Si, and H. Chen, "Document-level relation extraction as semantic segmentation," *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed., pp. 3999–4006, 8 2021, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/551

[15] F. Christopoulou, M. Miwa, and S. Ananiadou, "Connecting the dots: Document-level neural relation extraction with edge-oriented graphs," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4925–4936, Nov. 2019. [Online]. Available: https://aclanthology.org/D19-1498

[16] F. Wang, F. Li, H. Fei, J. Li, S. Wu, F. Su, W. Shi, D. Ji, and B. Cai, "Entity-centered cross-document relation extraction," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9871–9881, Dec. 2022. [Online]. Available: https://aclanthology.org/2022.emnlp-main.671

[17] 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.

[18] 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

[19] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, Vol. 7, No. 3, pp. 535–547, 2019.

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," *Proceedings*

of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45, Oct. 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6