문서 요약 데이터셋을 이용한 생성형 근거 추론 방법

장예진O, 장영진, 김학수
건국대학교 인공지능학과
dpwls258@konkuk.ac.krO, danyon@konkuk.ac.kr, nlpdrkim@konkuk.ac.kr

Generative Evidence Inference Method using Document Summarization Dataset

Yeajin JangO, Youngjin Jang, Harksoo Kim
Department of Artificial Intelligence, Konkuk University

요 약

자연어처리는 인공지능 발전과 함께 주목받는 분야로 컴퓨터가 인간의 언어를 이해하게 하는 기술이다. 그러나 많은 인공지능 모델은 블랙박스처럼 동작하여 그 원리를 해석하거나 이해하기 힘들다는 문제점이 있다. 이 문제를 해결하기 위해 설명 가능한 인공지능의 중요성이 강조되고 있으며, 활발히 연구되고 있다. 연구 초기에는 모델의 예측에 큰 영향을 끼치는 단어나 절을 근거로 추출했지만 문제 해결을 위한 단서 수준에 그쳤으며, 이후 문장 단위의 근거로 확장된 연구가 수행되었다. 하지만 문서 내에 서로 떨어져 있는 근거 문장 사이에 누락된 문맥 정보로 인하여 이해에 어려움을 줄 수 있다. 따라서 본 논문에서는 사람에게 보다 이해하기 쉬운 근거를 제공하기 위한 생성형 기반의 근거 추론 연구를 수행하고자 한다. 높은 수준의 자연어 이해 능력이 필요한 문서 요약 데이터셋을 활용하여 근거를 생성하고자 하며, 실험을 통해 일부 기계독해 데이터 샘플에서 예측에 대한 적절한 근거를 제공하는 것을 확인했다.

주제어: 설명 가능한 인공지능, 문서요약, 기계독해, 생성형 근거 추론

1. 서론

자연어처리(Natura Language Processing; NLP)는 컴퓨터가 인간의 언어를 이해하고 처리할 수 있도록 연구하는 분야로 언어의 복잡성과 다양성을 파악하고 처리하는데 중점을 두고 발전해왔다. 최근에는 인공지능(Artificial Intelligence; AI)의 도움을 받아 다양한 분야에서 높은 성능을 달성하고 있다. 특히, 자연어처리뿐만 아니라 이미지 및 음성 인식 분야에서도 인공지능의 성능이 인간을 뛰어넘는 경우가 다수 관측된다[1]. 그러나 기존의 인공지능 모델은 대부분 블랙박스와 같이 작동 원리나 결정 근거가 불투명한 문제점을 가지고 있다[2]. 이러한 특성 때문에 모델의 예측이나 결정에 대한 신뢰성에 대한 문제가 지속해서 제기되었다[3,4]. 이를 해결하기 위해 설명 가능한 인공지능(XAI; eXplainable Artificial Intelligence) 연구의 중요성이 높아졌다[4]. 설명 가능한 인공지능은 모델의 예측이나 결정 과정을 이해할 수 있도록 도와주며, 이를 통해 사용자나 연구자가 모델 작동 원리나 근거를 명확하게 파악할 수 있도록 한다. 예를 들면 의료 분야와 같은 전문지식 분야에서 인공지능 모델은 신뢰성을 높이기 위해 진단 뿐만 아니라 이에 대한 근거를 제공해야한다[4]. 이렇게 설명 가능한 인공지능은 모델의 예측이나 결정에 대한 신뢰성을 높이고, 잘못된 예측이나 결정에 대한 원인 분석도 용이하게 만든다.

따라서 본 논문에서는 위의 설명 가능한 인공지능의 중요성을 바탕으로 문서 요약 데이터셋을 활용하여 기계독해(Machine Reading Conprehension; MRC) 작업의 정답 추출에 대한 근거를 생성하는 방법론을 제안한다.

2. 관련 연구

최근 자연어 추론(Natural Language Inference; NLI)과 기계독해 분야에서는 모델의 예측을 설명하거나 근거를 제공해주는 연구가 활발하게 이루어지고 있다. 특히 [5]에서는 자연어 추론 태스크에서 alignment-based 접근법을 사용하여 모델의 예측을 설명하고자 했다. 그러나 위 연구는 단어나 구 단위의 특징을 중심으로 설명을 제공하기 때문에 예측에 대한 근거보다는 문제 해결에 대한 단서 수준에 그쳤다. 또한 [6]에서는 의료 분야에서 예측에 대한 근거를 추출하기 위해 다양한 수준의 텍스트 특징(단어나 절)을 사용하여 설명을 제공했다. 그러나 이 연구에서도 짧은 단어나 절 단위 근거를 제공하였기 때문에 문맥적인 연결성이 무시되는 문제가 있었다. 기존의 연구에서는 주로 단어나 문장 단위의 근거를 추출하는 방식을 취했다. 그러나 이러한 방법들은 문맥을 완전히 반영하지 못하거나, 해석하기 어려운 문제점이 있다. 따라서 우리는 본 논문을 통해 사용자에게 보다 직관적이고 명확한 근거를 제시하고자 하며, 이를 통해서 모델의 예측에 대한 신뢰성을 더욱 높이고자 한다.

3. 제안 방법

본 논문에서는 모델 예측에 대한 근거를 자동으로 생성하기 위해 기계독해 데이터셋 학습 전, 문서 요약 데이터셋을 사전 학습한다. 이때 사용된 문서 요약 데이터셋은 논문 요약 데이터셋[9] 이다. 논문 요약 데이터셋은 논문 제목, 논문 초록, 논문 초록 요약 등의 메타 정

보를 포함하고 있다. 본 논문에서는 논문 제목을 쿼리라고 가정하여, 쿼리에 대한 논문 초록을 요약하고자 했다. 트랜스포머 인코더-디코더 구조를 가지는 한국어 BART를 초기 가중치로 설정했다.

3.2 문서 요약 사전학습

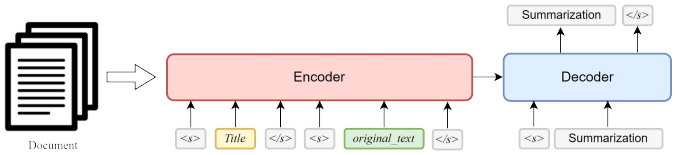

그림 1. 논문 데이터셋을 활용한 사전학습 구조도

위의 그림 1은 기계독해 데이터셋 학습 전, 문서 요약 데이터셋을 학습하는 과정을 보여준다. 토큰화된 논문 제목과 논문 초록을 <S> 토큰과 </S> 토큰으로 구분하여 하나의 입력으로 구성한다. 구성된 입력은 인코더에 입력되며 디코더를 통해 논문 초록 요약을 생성하도록 학습한다. 이는 논문의 초록을 기반으로 논문 제목과 밀접하게 연관된 논문 초록 요약을 생성하게 된다. 우리는 위 과정이 문서에서 질문에 대한 답을 찾아가는 기계독해 작업과 유사하다고 판단했으며, 디코더를 통해 생성된 요약문이 기계독해 작업에서 근거로 활용될 수 있을 것이라 판단했다. 위 과정에서 모델은 제목과 본문 간의 연관성을 깊게 파악하며 문맥을 반영한 요약문을 생성하도록 파인 튜닝된다.

3.3 기계독해 파인튜닝

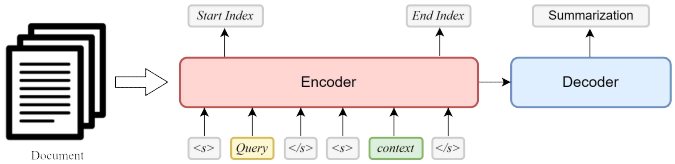

그림 2. MRC 데이터셋을 활용한 근거 생성 구조도

본 논문에서는 논문 요약 데이터셋을 사전 학습한 모델 가중치를 초기 값으로 설정하여 기계독해 데이터셋을 학습한다. 기계독해 데이터셋 학습 단계는 위의 그림 2와 같다. 모델의 입력은 토큰화된 질문과 문서를 <S> 토큰과 </S> 토큰으로 구분하여 하나의 입력으로 구성한다. 기계독해의 답변은 인코더를 통해 학습되며, 디코더는 초기 가중치(논문 요약 데이터셋 학습한 가중치)로 생성한 요약문을 학습한다. 이때 디코더가 학습하는 요약문은 기계독해 데이터 입력에 대해 생성된 요약문을 학습한다.

4. 실험

4.1 실험 준비

본 논문에서 논문 요약 데이터셋과 KorQuAD 1.0을 실험에 사용했다. 논문 요약 데이터셋은 282,580쌍으로 논문 제목, 논문 초록, 논문 초록 요약 등의 메타데이터를 포함한다. KorQuAD 1.0은 60,407쌍의 학습데이터와 5,774 쌍으로 이루어져 있다. KorQuAD 1.0에는 정답 추론에 필요한 근거가 부착되어있지 않기 때문에 성능 측정은 정성평가를 기준으로 계산했다.

4.2 정성 평가

본 논문에서 수행한 정성평가에는 KorQuAD 1.0 검증 데이터 중 30개의 샘플을 추출하여 질문-생성 근거-정답 쌍으로 정성평가를 진행했다. 평가 지표는 '근거 평가', '문법 평가', '사용자 이해도 평가'의 세가지 기준으로 나누었으며, 자세한 내용은 아래와 같다. 첫 번째로 '근거 평가'는 기계독해 데이터셋의 질문과 생성된 요약문을 바탕으로 해당 요약문만 보고도 정답을 도출할 수 있는지를 평가하는 기준이다. 이를 위해 0/0.5/1점의 척도로 평가하였다. 두 번째 평가 기준은 '문법 평가'이다. 이는 생성된 요약문이 언어적으로 올바르게 구성되어 있는지 즉 문법적으로 옳은지를 평가하는 기준이다. 이 평가는 0/1점의 척도로 진행되었다. 마지막으로 '사용자 이해도 평가'는 생성된 요약문의 문맥적인 흐름과 내용이 자연스럽고 이해하기 쉬운지를 평가하는 기준이다. 이를 위해 0/0.5/1점의 척도로 평가하였다. 총 3명의 평가자가 위의 세 가지 기준으로 평가를 진행하였으며 각 기준별로 얻은 점수의 평균을 표 1에 정리하였다.

표 1. 정성 평가 결과

	근거 평가	문법 평가	사용자 이해도 평가
평가자 1	0.400	0.933	0.817
평가자 2	0.300	0.900	0.883
평가자 3	0.367	0.800	0.717
평가자 평균	0.356	0.878	0.806

위의 표 1에 따르면 근거 평가와 비교하여 문법 평가 점수와 사용자 이해도 평가 점수가 상대적으로 높은 것을 알 수 있다. 이는 본 논문에서 사용한 사전학습모델에 내재되어 있던 지식으로 나온 결과로 판단된다. 이와 달리 근거 평가 점수의 경우 학습에 사용했었던 논문 요약 작업과 기계독해 작업의 차이로 인해서 낮은 점수가 측정된 것으로 판단된다. 우리는 모델의 생성 근거를 확인했다. 이는 아래의 표 2와 같다.

표 2. 정답에 대한 근거문 생성 예시

질문	정답	생성된 요약문
1985년 파리의 세계 실내 선수권 대회에서 자신의 최고 기록 2.13m로 은메달을 획득한 사람은?	소토마요르	소토마요르는 처음으로 2m를 제거할 때 15세 밖에 안 되었고, 1983년 말에 2.15m의 최고 기록을 세웠다
김이수가 나온 고등학교의 이름은?	전남고등학교	1953년 3월 24일 전라북도 고창군에서 태어난 김이수는 고수국민학교, 광주서중학교, 전남고등학교와 서울대학교 법학과를 졸업했다.
세운상가는 어떤 용도의 건물인가?	주상복합아파트	세운상가는 세계의 기운을 가진다는 뜻으로, 일제강점기에 공습을 대비해 소개(疏開)된 지역에 광복 이후 불법 판자촌이 들끓는 것을 보고 당시 불도저시장으로 불리던 김현옥 시장이 주도해 김수근 건축가가 설계한 대한민국 최초의 주상복합아파트가 탄생했다.

위 표 2의 첫 번째 예시에서는 질문에 대한 정답이 포함된 요약문이 생성되었으나 hallucination 현상이 발생한 것을 확인할 수 있다. 그러나 두 번째와 세 번째 예시에서는 문법적으로 완벽하며 문맥적으로도 자연스러운 요약문이 생성된 것을 알 수 있다. 이는 제안 모델에서도 초거대 인공지능 모델에서 나타나는 hallucination문제가 있음을 확인할 수 있었다.

5. 결론 및 향후 연구

본 연구에서는 문서 요약 데이터셋을 이용해 사전에 학습된 모델을 통하여 근거 요약문을 생성하는 방법론을 제안하였다. 해당 방법론의 정성평가 결과를 보면 사전 학습된 모델을 사용했기 때문에 문법 평가와 사용자 이해도 평가 면에서는 점수가 높은 것을 볼 수 있다. 그러나 근거 평가 점수가 다른 평가 점수들보다 낮음은 논문 데이터셋과 기계독해 데이터셋 사이에 존재하는 차이점 때문인 것으로 판단된다. 이러한 데이터 성격의 차이는 실험 결과에 영향을 미쳤을 가능성이 있으며 이는 본 연구의 한계점으로 볼 수 있다. 하지만 표 2에서 확인할 수 있듯, 일부 평가 데이터셋에서 올바르게 정답 추론에 필요한 근거를 생성하는 점을 알 수 있다. 따라서 우리는 단순히 생성된 요약문을 그대로 학습하는 것이 아닌 근거에 적합한 요약문을 학습할 수 있도록 향후 연구를 수행하고자 한다. 또한 인코더와 디코더의 가중치 업데이트가 기계독해 정답 추출에 미치는 영향을 깊게 분석하고 이를 개선하는 방향으로 연구를 진행할 계획이다. 또한 설명 가능한 인공지능 관점에서 볼 때 모델의 내부 동작 원리와 결정 근거를 더욱 투명하게 만드는 방법에 대한 연구도 필요하다. 이를 통해 사용자는 모델의 예측과 근거 생성 과정을 더욱 명확하게 이해할 수 있을것이며 이는 모델의 신뢰성을 높이는데 기여할 것으로 기대된다.

감사의 글

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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),pp. 770-778,2016

[2] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis,B. Kawas, and P. Sen, “A survey of the state of explainable AI for natural language processing,” Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 447-459, Dec.2020.

[3] M. Ribeiro, S. Singh, and C. Guestrin, “ “why shouldI trust you?”: Explaining the predictions of any classifier,” Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 97-101,Jun. 2016.

[4] Rai, A. Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science, 48(1), 137-141.2020

[5] Jain, D. K., Rahate, A., Joshi, G., Walambe, R.,

& Kotecha, K. Employing Co-Learning to Evaluate the Explainability of Multimodal Sentiment Analysis. IEEE Transactions on Computational Social Systems.2022
[6] H. Ngai and F. Rudzicz, "Doctor XAvIer: Explainable diagnosis on physician-patient dialogues and XAI evaluation," Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 337-344, May 2022.
[7] Z. Jiang, Y. Zhang, Z. Yang, J. Zhao, and K. Liu,"Alignment rationale for natural language inference,"Proceedings of the 59th Annual Meeting of the ssociation for Computational Linguistics and the 11th nternational Joint Conference on Natural Language rocessing (Volume 1: Long Papers), pp. 5372-5387, Aug. 2021.
[8] H. Wu, W. Chen, S. Xu, and B. Xu, "Counterfactualsupporting facts extraction for explainable medical record based diagnosis with graph network," Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1942-1955, Jun.2021.
[9]https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=90
[10] https://github.com/SKT-AI/KoBART