토픽 기반의 지식그래프를 이용한 BERT 모델

민찬욱[1], 안진현[2], 임동혁[3]

[1]광운대학교 인공지능융합학과
[2]제주대학교 경영정보학과
[3]광운대학교 정보융합학부
a4073631@kw.ac.kr, jha@jejunu.ac.kr, dhim@kw.ac.kr

Topic-based Knowledge Graph-BERT

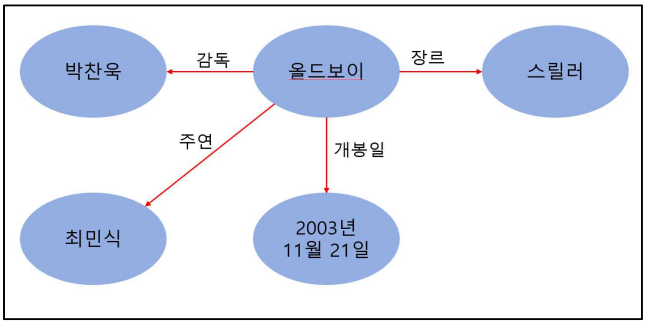
Chan-Wook Min[1], Jin-Hyun Ahn[2], Dong-Hyuk Im[3]
[1]Dept. of Applied Artificial Intelligence, Kwang-Woon University
[2]Dept. of Management Information System, Jeju University
[3]School of Information Convergence, Kwang-Woon University

요 약

최근 딥러닝의 기술발전으로 자연어 처리 분야에서 Q&A, 문장추천, 개체명 인식 등 다양한 연구가 진행 되고 있다. 딥러닝 기반 자연어 처리에서 좋은 성능을 보이는 트랜스포머 기반 BERT 모델의 성능향상에 대한 다양한 연구도 함께 진행되고 있다. 본 논문에서는 토픽모델인 잠재 디리클레 할당을 이용한 토픽별 지식그래프 분류와 입력문장의 토픽을 추론하는 방법으로 K-BERT 모델을 학습한다. 분류된 토픽 지식그래프와 추론된 토픽을 이용해 K-BERT 모델에서 대용량 지식그래프 사용의 효율적 방법을 제안한다.

1. 서론

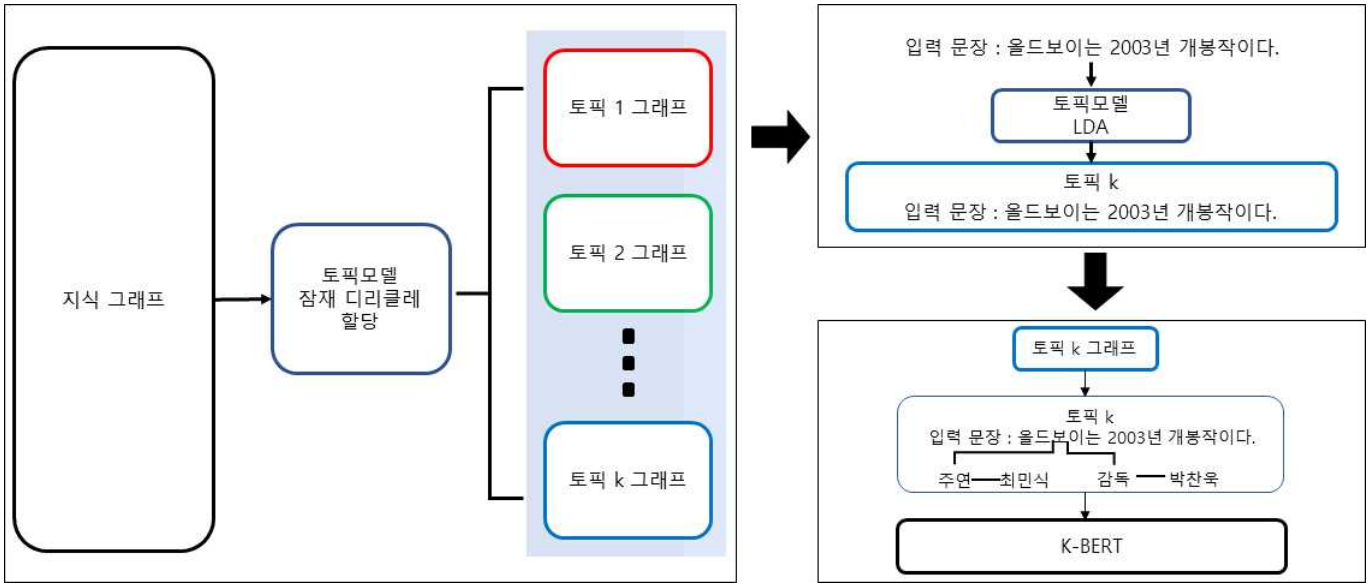
최근 딥러닝 기술의 발전으로 자연어에 대한 다양한 연구가 진행되고 있다[1]. 자연어 처리는 머신러닝을 적용한 Q&A, 문장추천, 기계번역, 개체명인식, 감성분석 등의 분야가 있다[2]. 딥러닝 기반 자연어 처리에서는 트랜스포머기반의 BERT 모델이 다양한 자연어 처리 분야에서 높은 성능을 보이고 있다[3]. BERT 모델은 대용량의 텍스트 데이터를 이용해 사전학습 모델을 생성한다. BERT 모델로 생성된 사전학습모델에 전이학습을 적용해 다양한 분야에서 높은 성능을 보이는 자연어 처리 모델을 생성할 수 있다. 자연어 처리에 대한 연구가 활발해짐에 따라 정확한 문맥파악을 위한 지식 그래프, 토픽 모델링의 관심도 높아지고 있다[4, 5]. 지식 그래프는 (그림 1)의 예시처럼 단어들 사이의 관계를 그래프로 나타낸 것으로 <주어, 술어, 목적어> 형식의 트리플 구조로 표현된다. 지식그래프를 자연어 처리 모델에 적용해 부족한 지식을 추가로 학습할 수 있다[6]. 하지만 방대한 양의 지식그래프를 사용 시 데이터에 혼란을 발생시킬 수 있는 지식노이즈 현상과 대용량의 지식그래프를 탐색하는 단점이 존재한다. 본 논문에서는 이러한 현상을 개선하기 위해 지식그래프를 토픽모델링을 이용해 토픽별로 나누고 입력 데이터의 토픽을 파악해 해당 토픽에 일치하는 지식을 추가하는 Topic-based Knowledge Graph-BERT 모델을 제안한다. 제안하는 모델은 토픽별 지식추가를 진행하기 때문에 입력 문장의 토픽과 다른 지식을 추가하는 것을 방지할 수 있다.



(그림 1) 지식그래프 예시.

2. 잠재 디리클레 할당

잠재 디리클레 할당이란 토픽 모델링의 기법중 하나이다[7]. 문서들은 토픽들의 혼합으로 구성되어져 있으며, 토픽들은 확률 분포에 기반하여 단어들



(그림 2) Topic-based Knowledge Graph-BERT.

을 생성한다고 가정한다. 가정을 이용해 데이터의 생성 과정을 역추적해 문서들의 토픽 분포를 도출하는 기법이다. (1)의 수식은 D는 말뭉치 전체의 수, K는 토픽의 수, N은 d번째 문서의 단어수를 의미한다. 정리하면 d번째 문서 i번째 단어의 토픽 $z_{d,i}$가 j번째에 할당될 확률을 나타낸다.

$$p(z_{d,i}=j|z_{-i},w)=\frac{n_{d,k}+\alpha_j}{\sum_{i=1}^{K}(n_{d,i}+\alpha_i)}\times\frac{v_{k,w_{d,n}}+\beta_{w_{d,n}}}{\sum_{j=1}^{V}(v_{k,j}+\beta_j)} \quad (1)$$

본 논문에서는 잠재 디리클레 할당을 이용해 지식그래프를 토픽별로 분류하고 토픽모델을 생성한다.

3. Topic-based Knowledge Graph-BERT

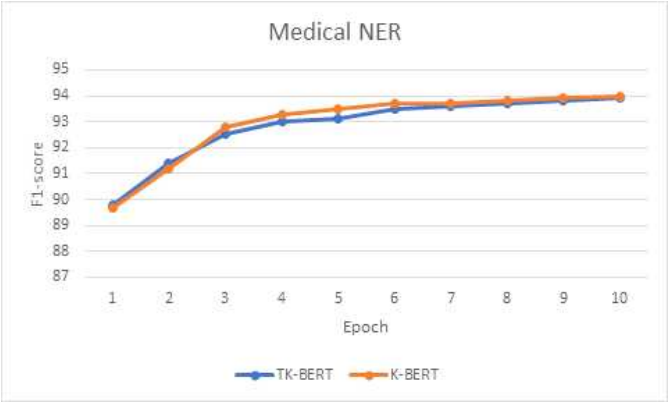
Topic-based Knowledge Graph-BERT는 토픽별 지식그래프 분류 단계와 입력문장의 토픽을 추론하는 토픽 추론 단계, 토픽에 알맞은 지식을 추가하는 지식 추가단계로 구성되어 있다. (그림 2)는 Topic-based Knowledge Graph-BERT 모델의 전반적인 과정을 나타낸다. 먼저 대용량의 지식그래프를 토픽모델인 잠재 디리클레 할당을 이용해 토픽에 맞게 분류한다. 다음으로 지식그래프로 생성된 토픽모델을 이용해 입력문장의 토픽을 도출한다. 마지막으로 도출된 토픽과 일치하는 지식그래프의 지식을 추가해 K-BERT모델로 전이 학습을 진행한다.

4. 실험

실험에 사용한 데이터 셋은 의료 개체명인식 데이터와 의료 지식그래프를 이용했다. 의료 개체명 인식 데이터는 문장으로 구성되어 있고 문장을 구성하는 각 단어들이 어떤 증상인지 라벨링 되어 있다. 의료 지식그래프는 잠재 디리클레 할당을 이용해 토픽별로 분류했다. 토픽별로 분류된 지식그래프를 이용해 지식 추가를 진행할 때 대용량의 지식 그래프가 아닌 토픽으로 분류된 지식그래프를 활용해 지식 추가를 진행하였다. 모델평가는 토큰 단위로 F1-score를 이용해 진행했다. 실험 결과(그림 3)는 대용량의 지식그래프를 이용해 지식을 추가한 K-BERT 모델과 문장의 토픽을 추론해 해당 토픽의 지식그래프의 지식만 추가한 TK-BERT 모델의 F1-score가 크게 차이가 나지 않는 것을 볼 수 있었다.

5. 결론

본 논문은 지식그래프를 이용해 지식을 추가 학



(그림 3) 실험 결과.

습하는 K-BERT 모델을 기반으로 토픽별 지식그래프 분류, 학습하는 모델을 제안하였다. 제안한 모델은 대용량 지식그래프를 토픽별로 분류하고 입력문장에 대해 토픽을 추론, 토픽에 맞는 지식을 추가해 학습하는 방법이다. 이는 대용량의 지식그래프의 크기를 줄여 학습 속도 향상에 기여하고 대용량의 지식그래프를 학습한 K-BERT 모델과의 성능차이도 크게 나지 않는 것을 확인했다. 지식그래프를 토픽 모델링 기반으로 분류해 효율적으로 사용하는 TK-BERT 모델을 제안한다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2021R1F1A1054739).

또한, 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2022-2018-0-01417).

참고문헌

[1] CHIU, Billy, et al. "How to train good word embeddings for biomedical NLP." Proceedings of the 15th workshop on biomedical natural language processing. 2016. p. 166-174.

[2] NA, Hyung-Sun, et al. "Sentence Recommendation Using Beam Search in a Military Intelligent Image Analysis System." KIPS Transactions on Software and Data Engineering, 2021, 10.11: 521-528.

[3] DEVLIN, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. p. 4171-4186.

[4] KIM, Seongyong, et al. "Semantic Scene Graph Generation Using RDF Model and Deep Learning." Applied Sciences, 2021, 11.2: 826.

[5] AIT-MLOUK, Addi; JIANG, Lili. "KBot: a Knowledge graph based chatBot for natural language understanding over linked data." IEEE Access, 2020, 8: 149220-149230.

[6] LIU, Weijie, et al. "K-bert: Enabling language representation with knowledge graph." Proceedings of the AAAI Conference on Artificial Intelligence. 2020. p. 2901-2908.

[7] BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. "Latent dirichlet allocation." Journal of machine Learning research, 2003, 3.Jan: 993-1022.