대규모 지식그래프와 딥러닝 언어모델을 활용한 기계 독해 기술

김성현*, 김성만*, 황석현*
*㈜솔트룩스, AI Labs
e-mail : seonghyunkim@saltlux.com

Machine Reading Comprehension based on Language Model with Knowledge Graph

Seonghyun Kim *, Sungman Kim*, Seokhyun Hwang*
*AI Labs, Saltlux Inc.

요 약

기계 독해 기술은 기계가 주어진 비정형 문서 내에서 사용자의 질문을 이해하여 답변을 하는 기술로써, 챗봇이나 스마트 스피커 등, 사용자 질의응답 분야에서 핵심이 되는 기술 중 하나이다. 최근 딥러닝을 이용한 기학습 언어모델과 전이학습을 통해 사람의 기계 독해 능력을 뛰어넘는 방법론들이 제시되었다. 하지만 이러한 방식은 사람이 인식하는 질의응답 방법과 달리, 개체가 가지는 의미론(Semantic) 관점보다는 토큰 단위로 분리된 개체의 형태(Syntactic)와 등장하는 문맥(Context)에 의존해 기계 독해를 수행하였다. 본 논문에서는 기존의 높은 성능을 나타내던 기학습 언어모델에 대규모 지식그래프에 등장하는 개체 정보를 함께 학습함으로써, 의미학적 정보를 반영하는 방법을 제시한다. 본 논문이 제시하는 방법을 통해 기존 방법보다 기계 독해 분야에서 높은 성능향상 결과를 얻을 수 있었다.

1. 서론

질의응답 시스템은 사용자의 질문에 따른 답변을 만들어내는 기술로, 챗봇이나 스마트 스피커 등의 핵심 기술로 여겨지고 있다. 기존의 질의응답 시스템은 정형화된 지식 그래프(Knowledge graph, KG)를 구축하여 질의를 처리하는 지식 기반(Knowledge base, KB) 시스템과, 비정형 데이터에서의 질의응답을 수행할 수 있는 정보 추출(Information retrieval, IR) 기반의 시스템으로 구분할 수 있다. KB 질의응답 시스템은 질의응답의 대상이 되는 개체(Entity) 간의 지식 정보를 트리플(Triple) 구조인 주어부(Subject)-서술부(Predicate)-목적어부(Object)로 저장한 KG 를 활용한다 [1,2]. 따라서 IR 기반의 질의응답 시스템에 비해 상대적으로 정확한 답변을 내놓을 수 있다는 장점이 있지만, KG 에 존재하지 않는 개체에 관한 질의응답을 수행할 수 없다는 단점이 있다. 반면 IR 기반의 질의응답 시스템은 KG 를 구축하고 유지할 필요가 없으며, 비정형의 문서와 사전에 정의하지 않은 개체에 대해서도 답변을 수행할 수 있다는 장점이 있다 [3, 4]. 이러한 IR 기반 질의응답 시스템의 기반 기술로 기계 독해(Machine reading comprehension, MRC)가 있다 [5]. MRC 란, 사용자의 질문을 받은 기계가 관련된 문서를 읽고 이해하여 문서 내에서 질의에 해당하는 정확한 답변을 찾아내는 기술을 말한다. 딥러닝 기술이 빠르게 발전하면서, 최근에는 방대한 데이터를 이용해 기학습된 언어모델과 전이학습을 통해 자연어를 처리하는 다양한 방법들이 소개되어지고 있다. 특히, Google 의 BERT (Bi-directional Encoder Representations from Transformers) 모델은 다양한 자연어 처리 분야의 챌린지에서 가장 좋은 성적을 내면서 주목받았다 [6]. BERT 는 영어 MRC 데이터셋인 SQuAD [7]에서 등장 즉시 가장 높은 성적을 보였고, 영어권뿐만 아니라 한국어 MRC 데이터셋인 KorQuAD 에서도 가장 높은 성적을 나타냈다 [8, 9]. 하지만 이러한 방식은 사람이 개체에 대해 가지고 있는 지식을 활용에 질의응답을 수행하는 방식과 달리, 개체가 가지는 의미론(Semantic) 관점 보다는 개체의 형태(Syntactic)와 그 개체가 등장하는 문맥(Context)에 의존하여 MRC 를 수행한다. 본 논문에서는 KB 의 정보를 BERT 모델의 학습자질로 사용함으로써 KB 기반의 질의응답과 IR 기반의 질의응답을 융합한, 더 좋은 성능의 MRC 기술 방법론을 제시하고자 한다.

2. KG 와 언어모델을 활용한 MRC 모델

이번 장에서는 KG 와 언어모델을 활용한 MRC 모델을 구축하는 방법에 대해서 소개한다. 본 논문에서는 KG 의 지식 정보를 BERT 언어모델 학습의 학습자질로 포함시킴으로써 개체에 대한 의미를 MRC 과

정에 반영하고자 한다. KG 는 공개된 API[1]를 사용하였으며, <표 1>과 같이 총 40 개의 서술부로 지식을 제한하였다. 이로써 총 707,667 개의 개체에 대한 2,214,641 개의 트리플 데이터를 추출할 수 있었다.

<표 1> 언어모델 학습에 사용된 서술부의 종류

서술부 명칭	의미
leader	대표, 지도자
parent	부모
bornIn	출생지
capital	수도
education	모교, 출신학교
activity	가담하다, 활동하다
artist	가수, 작가
buriedIn	묻히다, 안장지
channel	방송채널
ideology	사상, 이념, 정치노선
majorWork	대표작, 주요작품
language	모국어
relatedLocation	관련지역, 인접지역
type	유형, 종류
spouse	배우자
influence	영향을 끼치다
commander	지휘관
genre	장르, 유형
member	구성원, 회원
employer	근무기관, 근무지
influencedBy	영향을 받다
happenedIn	발생지
player	연기자
owner	소유자, 주인
creator	개발자, 발명자
manager	소속사, 관리자, 운영자
director	감독, 연출자
composer	작곡가
distributor	배급업자, 유통업자
religion	종교
diedIn	사망지
nationality	국적
field	분야, 종목
locatedIn	위치, 장소, 주소
producer	제작자
job	직업
user	사용자
mainlyLocatedIn	본교위치, 본사위치
child	자녀, 자식
relative	친족, 친척

학습 데이터로 2019 년 06 월 20 일에 공개된 Kowiki[2] 덤프 데이터를 활용하였다. 크기는 약 570M 로, 3,772,605 문장, 48,739,306 어절로 구성되어 있다.

2.1 데이터 전처리

BERT 언어모델 학습에 KG 정보를 학습 자질로써 사용하기 위해, <표 2>에서 제시된 방식에 따라 학습 데이터 전처리를 수행하였다.

<표 2> 언어모델 학습을 위한 전처리 과정의 예시

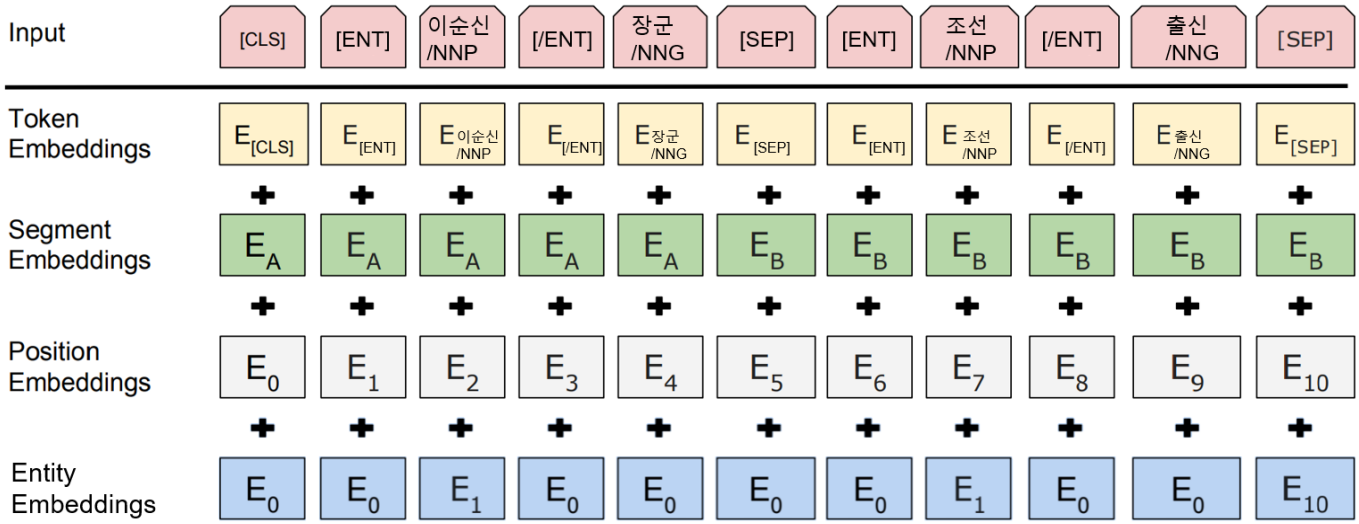
단계	예시
원본 문장	이순신은 조선 중기의 무신이다.
주요 개체 후보 추출	이순신, 조선, 조선 중기
주요 개체 추출	이순신, 조선
개체명 테깅	[ENT] 이순신 [/ENT] 은 [ENT] 조선 [/ENT] 중기의 무신이다.
형태소 테깅	이순신/NNP 은/JX 조선/NNP 중기/NNG 의/JKG 무신/NNG 이/VCP 다/EP ./SF
전처리 완료	[ENT] 이순신/NNP [/ENT] 은/JX [ENT] 조선/NNP [/ENT] 중기/NNG 의/JKG 무신/NNG 이/VCP 다/EP ./SF

먼저, 원본 문장에서 표면형 문자열 매칭을 통해 KG 에 존재하는 주요 개체 후보를 찾아낸다. 주요 개체 후보들을 각각 트리플 구조의 주어부와 서술어부로 조합(Combination)을 수행하고, KG 에 해당 개체들의 트리플이 존재할 경우 주요 개체로 추출한다. 추출된 주요 개체들을 이용해 원본 문장에서 개체명 테깅을 수행한다. 이 때, 개체의 시작에는 [ENT] 테그를, 개체의 끝에는 [/ENT] 테그를 부착함으로써 개체의 시작과 끝을 표현한다. 다음으로 원본 문장을 대상으로 형분석을 수행한 후, 형태소 테그를 부착해준다. 여기서 형분석기는 공개된 API 를 사용하였다. 최종적으로 전처리가 완료된 문장을 학습 데이터의 입력으로 사용하였다.

2.2 언어모델의 학습

KG 에서 획득한 개체에 대한 정보를 BERT 학습에 반영하기 위해, (그림 1)과 같이 언어모델 임베딩 층에 개체 임베딩 층(Entity embedding layer)을 추가하였다. 개체 임베딩 층은 형태소 단위로 잘려진 토큰이 개체 정보를 포함하고 있을 경우 1 로, 그렇지 않을 경우엔 0 으로 해당 토큰이 개체임을 표현한다. BERT 언어모델이 학습에 적용한 마스킹(Masking)이 해당 개체에 발생할 경우, 개체 테그([ENT], [/ENT]) 사이에 존재하는 모든 토큰을 함께 마스킹한다. 또한, 개체 테그와 고유명사 테그(NNP)가 부착된 토큰을 우선적으로 마스킹한다. 마스킹의 확률 및 마스킹 대상의 개수는 기존 BERT 와 동일하게 적용된다. 학습에 사용된 매

[1] https://www.adams.ai/
[2] https://dumps.wikimedia.org/kowiki/20190620/



(그림 1) $BERT_{Entity}$ 모델의 임베딩 레이어

개변수는 공개된 디폴트 값을 동일하게 따른다[3]. 학습은 Google TPU v2 를 사용하였으며, 128 배치 사이즈로 300,000 스텝을 학습하였다. 이는 Kowiki 덤프 데이터를 기준으로 약 10 epochs 정도에 해당하는 학습량이다. 최대 시퀀스 길이는 512 시퀀스로 학습하였다. 원시코드는 온라인 저장소[4]를 통해 공유 받을 수 있다.

3. 실험 및 성능평가 방법

제안하는 방법론과 비교를 위해 BERT 베이스라인 기학습 모델을 함께 구축하였으며, Google 에서 공개한 워드피스 토크나이징(Wordpiece tokenizing)을 이용해, 동일한 학습 데이터로 동일한 스텝만큼 학습을 수행하였다.

본 논문에서는 성능평가를 위해 기학습된 BERT 모델을 이용해 한국어 MRC 데이터셋인 KorQuAD v1.0[5]로 전이학습을 수행하고, 그 결과를 평가하였다. MRC 학습은 32 배치 사이즈, 3 epochs 로 학습을 하였으며, 최대 시퀀스 길이는 512, doc stride 는 256, 그리고 learning rate 는 3e-5 로 학습하였다. 성능 평가는 KorQuAD 의 dev set 을 기준으로 평가하였다. MRC 를 통해 출력된 결과가 데이터셋의 정답과 정확하게 일치하는지 여부를 결정하는 exact matching 점수와 정답과의 음절 유사도를 결정하는 F1 점수를 이용해 평가하였으며, KorQuAD 에서 공개한 평가코드를 사용하였다.

4. 실험 결과

<표 3>은 기존 BERT 의 학습 방식을 사용한 모델($BERT_{base}$)과 제안하고자 하는 모델($BERT_{Entity}$)을 KorQuAD MRC 데이터셋을 이용해 평가분석한 표이다. 정답과의 음절 유사도를 평가한 F1 의 경우, 베이스라인 모델 대비 4%가 향상되는 결과를 확인할 수 있었으며, 특히 정확도를 나타내는 exact matching 점수는 14 점이나 향상되는 결과를 얻을 수 있었다.

<표 3> KorQuAD MRC 성능 평가 결과

모델	Exact matching	F1
$BERT_{base}$	64.51	83.76
$BERT_{Entity}$	78.13	87.25

5. 결론 및 향후 연구

본 논문에서는 대규모 KG 와 BERT 언어모델을 이용해 MRC 문제를 해결하는 방법을 제안하였다. 언어모델과 MRC 의 학습 과정에서 개체에 대한 정보를 학습 자질로써 사용하는 방법을 통해 적은 데이터와 조금의 학습으로도 사람과 가까운 MRC 성능을 보여주었으며, 특히 기존 BERT 모델 대비 높은 성능 향상을 관찰할 수 있었다. 본 논문에서 제안하는 모델의 개체 임베딩은 개체의 여부를 1 과 0 을 통해 표현하였으나, 추후에는 개체명 분석을 통해 해당 개체의 특성을 표현하는 정보를 임베딩에 사용하면 더욱 높은 성능 향상을 기대할 수 있을 것이다.

Acknowledgement

이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2013-0-00109,WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] Cui, Wanyun, et al. "KBQA: learning question answering over QA corpora and knowledge bases." Proceedings of the VLDB Endowment 10.5 (2017): 565-576.

[2] Yih, Wen-tau, and Hao Ma. "Question answering with knowledge base, web and beyond." Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2016.

[3] https://github.com/google-research/bert
[4] https://github.com/MrBananaHuman/BertEntity
[5] https://korquad.github.io/category/1.0_KOR.html

[3] Chen, Danqi, et al. "Reading wikipedia to answer open-domain questions." arXiv preprint arXiv:1704.00051 (2017).
[4] Joshi, Mandar, et al. "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension." arXiv preprint arXiv:1705.03551 (2017).
[5] Ng, Hwee Tou, Leong Hwee Teo, and Jennifer Lai Pheng Kwan. "A machine learning approach to answering questions for reading comprehension tests." Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13. Association for Computational Linguistics, 2000.
[6] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[7] Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).
[8] 임승영, 김명지, and 이주열. "KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋." 한국정보과학회 학술발표논문집 (2018): 539-541.
[9] 박광현, et al. "BERT 와 Multi-level Co-Attention Fusion 을 이용한 한국어 기계독해." 한국정보과학회 학술발표논문집 (2019): 643-645.