기계 독해를 이용한 COVID-19 뉴스 도메인의 한국어 질의응답 챗봇

이태민†, 박기남†, 박정배†, 정영희‡, 채정민‡, 임희석†✉
고려대학교 Human-Inspired AI 연구소†, 버시스트‡
taeminlee@korea.ac.kr, spknn@korea.ac.kr, insmile@korea.ac.kr, icecoolof@gmail.com, onacloud@gmail.com, limhseok@korea.ac.kr

Korean Q&A Chatbot for COVID-19 News Domains Using Machine Reading Comprehension

Taemin Lee† , Kinam Park† , Jeongbae Park† , Younghee Jeong‡ , Jeongmin Chae‡ , Heuiseok Lim† ✉
Korea University Human-Inspired AI Research Institute† , Virssist‡

요 약

코로나 19와 관련한 다양한 정보 확인 욕구를 충족하기 위해 한국어 뉴스 데이터 기반의 질의응답 챗봇을 설계하고 구현하였다. BM25 기반의 문서 검색기, 사전 언어 모형인 KoBERT 기반의 문서 독해기, 정답 생성기의 세 가지 모듈을 중심으로 시스템을 설계하였다. 뉴스, 위키, 통계 정보를 수집하여 웹 기반의 챗봇 인터페이스로 질의응답이 가능하도록 구현하였다. 구현 결과는 http://demo.tmkor.com:36200/mrcv2 페이지에서 접근 및 사용을 할 수 있다.

주제어: MRC, chatbot, BERT, KorQuAD

1. 서론

코로나 19는 2019년 말 최초 발견된 이후 전 세계로 급격히 전염되어 각국에 심대한 손해를 끼치고 있다. 코로나 19로 야기되는 혼란은 이에 대한 정확한 정보의 습득으로 일부 해소가 가능하다. 하루가 다르게 지속해서 변화하고 증가하는 코로나 19 정보를 빠르게 습득하기 위해서는, 정보를 정제하고 사용자에게 즉시 전달하는 도구가 필요하다. 이에 본 연구진은 코로나 19의 정보를 MRC(Machine Reading Comprehension; 기계 독해) 기술을 이용하여 정제하고, 챗봇의 인터페이스로 전달하는 시스템을 설계하고 구현하였다.

2. 관련 연구

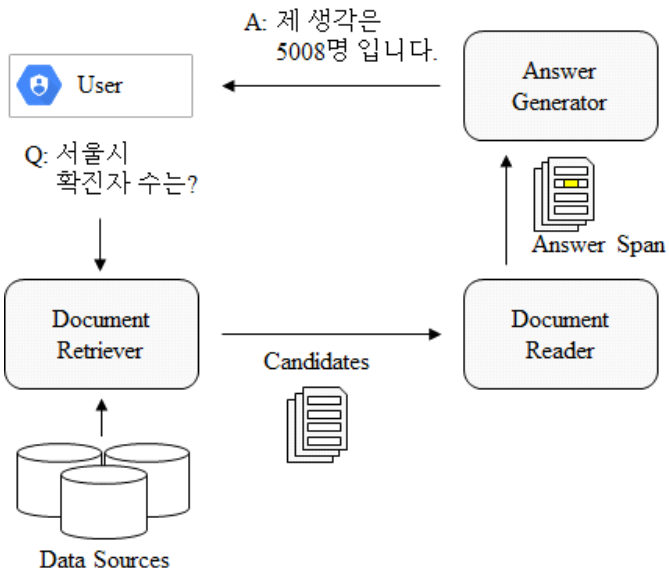
본 연구의 관련 연구는 PLM(Pretrained Language Model; 사전 학습 언어 모형)을 이용한 MRC 기법과 오픈 도메인에서의 MRC 기법의 두 가지로 나누어진다.

MRC 연구는 영어의 경우 SQuAD 데이터 집합[1], 한국어의 경우 KorQuAD 데이터 집합[2]을 이용하여 수행되고 있다. 해당 데이터 집합은 질문과 근거 문맥이 주어졌을 때 알맞은 근거 문맥 내 정답의 위치정보를 담고 있다. 정답의 위치정보를 알기 위해 토큰화된 문장을 인코딩하고 이를 이용해 근거 문맥 내 토큰의 정답 확률을 구하고 있다. 이러한 방식은 문맥 정보를 인코딩하기 위해 RNN(Recurrent Neural Network; 순환 신경망) 구조와 어텐션 메카니즘을 도입한 BiDAF[3] 연구에서 명확히 확인할 수 있다. MRC 연구의 최대 변곡점은 RNN 구조 대신 어텐션 메카니즘으로 신경망을 구축하는 Transformer 구조[4]의 발견이었다. 거대한 자연어 데이터 집합을 Transformer 구조의 인코딩 레이어로 학습한 BERT[5]는 단어와 문장에 대한 문맥이 반영된 인코딩을 가능하게 하였다. PLM인 BERT를 이용하여 MRC 연구가 수행되었으며, F1 점수 기준으로 사람 수준의 MRC 수준을 보여주었다. 이후 BERT 구조를 개선한 다양한 PLM을 이용한 MRC 연구가 진행되고 있다.

오픈 도메인의 MRC 연구는 근거 문맥과 질문의 쌍이 사전에 주어지지 않는 것이다. 질문에 알맞는 근거 문맥을 찾고, 그 안의 정답 위치정보를 찾는 문제를 다룬다. 즉, 일반적인 MRC와의 차이점은 정답이 포함될 검색 공간이 한정되지 않음에 있다. DrQA[6]는 위키피디아 문서 집합을 검색 공간으로 문제를 설정하였으며, TF-IDF를 이용하여 근거 문맥을 찾는 연구를 수행하였다. DenSPI[7]는 문맥과 질문이 유사한 경우 문맥과 질문의

밀집 벡터 표상의 내적값이 커지도록 학습한다. 즉, metric 학습을 수행하여 밀집 벡터 표상을 학습하고, 이를 TF-IDF로 구한 희소 벡터와 결합하여 근거 문맥을 찾는다. DPR[8]은 밀집 벡터와 희소 벡터를 분리하여 연구를 진행하였으며, 밀집 벡터 학습 방법을 개선하였다.

3. COVID-19 한국어 질의응답 챗봇



[그림 1] COVID-19 한국어 질의응답 챗봇 데이터 흐름도

COVID-19 한국어 질의응답 챗봇은 사용자의 질의에 대하여 답변을 제공하는 프로그램으로, [그림 1]과 같이 문서 검색기(Document Retriever), 문서 독해기(Document Reader), 정답 생성기(Answer Generator)의 3개의 모듈을 중심으로 설계하였다. 검색기와 독해기를 분리한 구조를 선택한 이유는 질의부터 대답까지의 시간을 최소화하기 위함이다.

문서 검색기는 사용자 질의를 이용하여 전체 검색 공간의 문서를 일정 수의 후보 문서 군으로 필터링하는 모듈이다. 문서 검색기는 지속해서 증가하는 다양한 데이터 집합에서 정답이 포함될 가능성이 큰 후보 문서를 신속하게 검색할 수 있어야 한다. 문서 독해기는 개별 문서에 대해 질문의 정답 확률이 높은 영역을 선택하는 모듈로, PLM 기반의 MRC 기법이 사용된다. 정답 생성기는 사용자에게 반환할 문장을 생성하는 모듈이다. 문서 검색기, 문서 독해기의 결과를 이용하여 생성한다. 챗봇 인터페이스를 이용하므로 빠른 인식이 가능하도록 짧고 명료한 형식으로 가공해야 한다.

세 가지 핵심 모듈은 다양한 방법으로 구현될 수 있으므로, 각각의 모듈의 의존성을 최소화하여 후속연구에서 쉽게 개선할 수 있도록 하였다. 또한, 본 연구는 가능한 대중적인 방법론을 사용하여 구현하였다.

문서 검색기는 Elastic Search를 이용한다. 문서 검색기의 경우 전처리(analysis) 과정에 따라 검색 결과가 달라지게 된다. 적절한 전처리 방법을 찾기 위해 KorQuAD 1.0 데이터 집합의 근거 문장(context)을 인덱싱하고 질문(question)을 질의문으로 검색하여 정답(answer)의 포함 여부를 기준으로 top-k 정확도를 측정하였다.

<표 1> 전처리 과정에 따른 문서 검색기 성능 (데이터 집합 KorQuAD 1.0)

	top-1 정확도	top-5 정확도
공백 분절	71.68%	84.19%
형태소 분석 + 명사 추출	**88.92%**	**97.26%**

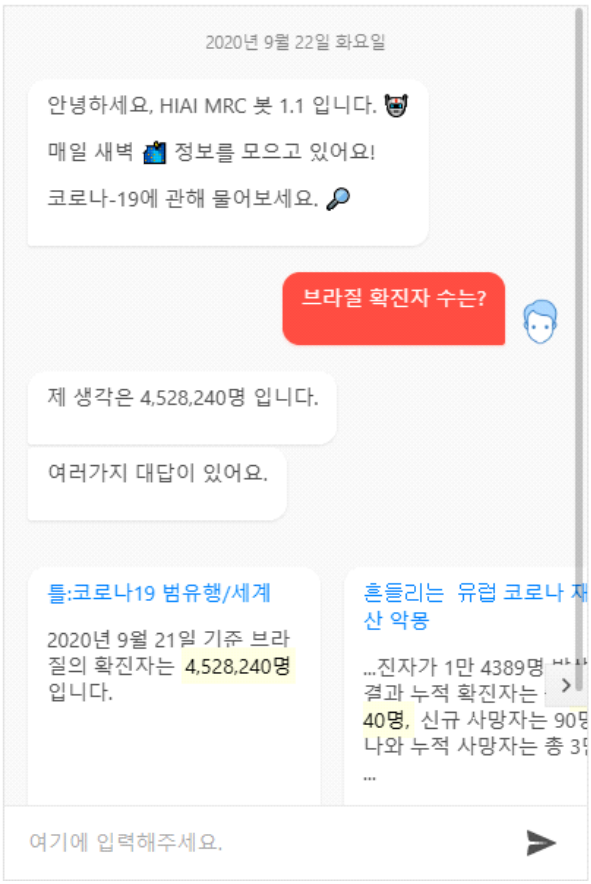
실험 결과의 경우 형태소 분석을 하여 명사를 추출하고 이를 이용하여 인덱싱할 경우 유의미한 성능 향상을 보였다. 질의가 들어오면 문서 집합에서 BM25 유사도가 높은 문서를 선택하며, 시간 경과로 유사도 점수를 decay 후보 문서 군을 반환한다.

정답 생성기는 BERT 구조를 이용하여 한국어 언어 모형을 학습한 PLM인 KoBERT[9]를 KorQuAD 데이터로 파인튜닝하여 구현하였다. 정답 생성기 모듈은 별도의 웹 서버로 동작하도록 하여 scale out이 가능하다. 정답 생성기는 문서 독해기의 softmax 확률값이 설정한 임계치를 넘을 경우 문장의 형식으로 가공하여 제공한다. 또한, 문서 검색기의 유사도 점수와 문서 독해기의 softmax 확률값을 곱한 값으로 순위화 한 후 상위 n개의 근거 문서를 카드 형식으로 가공하여 다양한 관점의 정보를 탐색할 수 있도록 하였다.

COVID-19 한국어 질의응답 챗봇의 질문과 응답의 예시는 다음 <표 2>와 같다.

<표 2> COVID-19 질의 응답 예시

Q	브라질 확진자 수는?
A	제 생각은 4,915,289명 입니다.
근거1	2020년 10월 6일 기준 브라질의 확진자는 **4,915,289명** 입니다.
근거2	...을 것이란 지적이 쏟아졌습니다. 결국 브라질은 확진자 **470만명**, 사망자 1만4천명을 훌쩍 뛰어넘으며 미국과 인도에 이...
Q	코로나 치료제는?
A	여러 가지 대답이 있어요.
근거1	...지다. 가장 주목을 받는 것은 리제네론사의 항체치료제 **REGN-COV2다.** 이 약물은 코로나 바이러스를 중화시키는 항체 두 가지...
근거2	...9) 치료제로도 활용되고 있는 에볼라 바이러스 치료제 **'렘데시비르'**, 인체면역결핍바이러스(HIV, 에이즈바이러스) 치료제 ...



[그림 2] COVID-19 한국어 질의응답 챗봇 화면

서비스를 위해 데이터 수집기와 챗봇 인터페이스를 구현하였으며, 데이터 수집기는 COVID-19의 정보를 수집하여 Elastic Search의 인덱스에 적재하도록 하였다. 데이터 수집기를 통해 COVID-19와 관련이 있는 뉴스 데이터, 위키피디아 문서, 국가별 누적 통계 정보를 수집하도록 하였다. 챗봇 인터페이스는 웹 프로그램으로 구현하여 PC, 스마트폰 등 다양한 기기에서 접근이 가능하다. 구현 결과는 http://demo.tmkor.com:36200/mrcv2 페이지에서 접근 및 사용을 할 수 있다.

4. 결론

코로나 19는 미시적, 거시적으로 대한민국에 큰 영향을 주고 있다. 이러한 상황에서 정확한 정보를 빠르게 인지하는 것은 혼란으로 야기되는 사회적 비용을 감소시키는 중요한 요인이다. 본 연구에서는 기계 독해 기술을 이용하여 COVID-19 뉴스에 대해 질의응답이 가능하도록 시스템을 설계하고 챗봇 인터페이스로 사용하도록 구현하였다. 코로나 19의 질의응답 챗봇은 개선의 여지가 있으며, 특히 검색기의 성능은 후속 작업에 영향을 미치므로 기계 독해에 적합한 검색 방법의 연구 개발이 필요할 것으로 보인다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과(IITP-2020-2018-0-01405) 및 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00368, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발)

참고문헌

[1] Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. “SQuAD: 100,000+ Questions for Machine Comprehension of Text.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1606.05250.

[2] 임승영, 김명지, 이주열. (2018). KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋. 한국정보과학회 학술발표논문집, (), 539-541.

[3] Seo, Minjoon, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. “Bidirectional Attention Flow for Machine Comprehension.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1611.01603.

[4] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. Ukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” In Advances in Neural Information Processing Systems 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5998-6008. Curran Associates, Inc.

[5] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1810.04805.

[6] Chen, Danqi, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. “Reading Wikipedia to Answer Open-Domain Questions.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1704.00051.

[7] Seo, Minjoon, Jinhyuk Lee, Tom Kwiatkowski, Ankur P. Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. “Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1906.05807.

[8] Karpukhin, Vladimir, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-Tau Yih. 2020. “Dense Passage Retrieval for Open-Domain Question Answering.” arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2004.04906.

[9] SKT brain, 2020 (accessed 2020.09.22). https://github.com/SKTBrain/KoBERT. I. Mani and T. Maybury, Advances in Automatic Text, The MIT Press, 1999.