BERT MRC를 활용한 한국 프로야구 Q&A 시스템

서정우○, 김창민, 김효진, 이현아
금오공과대학교 컴퓨터소프트웨어공학과
jungwoo7250@naver.com, min1003k@gmail.com, jing0318@naver.com, halee@kumoh.ac.kr

Korean Baseball League Q&A System Using BERT MRC

JungWoo Seo○, Changmin Kim, HyoJin Kim, Hyunah Lee
Kumoh National Institute of Technology, Dept. of Computer Software Engineering

요 약

매일 게시되는 다양한 프로야구 관련 기사에는 경기 결과, 각종 기록, 선수의 부상 등 다양한 정보가 뒤섞여있어, 사용자가 원하는 정보를 찾아내는 과정이 매우 번거롭다. 본 논문에서는 문서 검색과 기계 독해를 이용하여 야구 분야에 대한 Q&A 시스템을 제안한다. 기사를 형태소 분석하고 BM25 알고리즘으로 얻은 문서 가중치로 사용자 질의에 적합한 기사들을 선정하고 KorQuAD 1.0과 직접 구축한 프로야구 질의응답 데이터셋을 이용해 학습시킨 BERT 모델 기반 기계 독해로 답변 추출을 진행한다. 야구 특화 데이터셋을 추가하여 학습시켰을 때 F1 score, EM 모두 15% 내외의 정확도 향상을 보였다.

주제어: BERT, 프로야구, 기계 독해, 질의응답

1. 서론

한국 프로야구는 많은 사람들에게 사랑을 받고 있다. 그 인기를 증명하듯 경기 결과, 주요 기록, 부상 명단, 경기 리포트 등 수많은 야구 기사가 매일 쏟아져 나오고 있다. 코로나바이러스감염증-19로 인해 무관중 경기를 진행하는 상황에도 불구하고 꾸준히 경기가 진행되고 기사들이 작성되는 상황을 보아 프로야구 콘텐츠에 대한 수요가 지속될 것으로 보인다.

수많은 기사에서 직접 원하는 정보를 사용자가 일일이 찾는다는 것은 쉽지 않다. 한 명의 선수 이름으로 검색해도 수십 수백 개의 기사가 나오고, 그중 원하는 정보가 있을 기사를 선택해서 본문을 읽고 정보를 찾아내는 것은 굉장히 번거롭다. 카카오톡을 통해 서비스되는 '카카오톡 프로야구봇' 같은 정보 제공 챗봇 형태의 서비스는 정적인 입력에 정적인 답변을 하거나, 실시간으로 기록 혹은 뉴스 기사들을 제공하는 데에 그친다. 2018년 NC소프트에서 오픈한 페이지(PAIGE)[1]는 AI를 통한 야구 질의응답 기능을 제공하며, 여러 유형의 질문들에 대해 답변으로 좋은 평을 얻고 있다.

본 논문에서는 이처럼 프로야구에 대한 임의 질문에 답할 수 있는 시스템을 제안한다. 제안하는 시스템은 문서 선정과 정답 추출의 두 단계로 구성된다. 키워드를 기반으로 적합한 문서를 검색하는 연구들[2,3]에서는 키워드에 따라 차별화된 가중치를 부여하고 해당 가중치를 이용하여 문서에 점수를 매겨 순위화하여 검색한다. 본 논문에서는 이러한 접근을 사용하여 사용자의 질의에서 형태소 분석을 통해 키워드를 추출한 뒤 이를 이용해 적합한 한국 프로야구 기사 문서를 선정한다. 기사 적합도는 단어 가중치 기법인 BM25 알고리즘으로 계산한다.

정답 추출에서는 BERT를 이용한 기계독해를 사용한다. 2018년 Google에서 공개한 BERT의 등장으로 기계 독해 분야 성능이 크게 향상되었고, SQuAD[4], KorQuAD[5,6] 등의 기계 독해를 통한 질의응답 정확도 리더보드가 대부분 BERT 기반의 모델로 이루어져 있다. 본 시스템에서는 프로야구에 대한 용어나 기사를 학습시키기 위해 자체적으로 프로야구 질의응답 데이터 셋을 구축하여 학습에 활용한다.

2. 관련 연구

2.1 BM25 알고리즘

Okapi BM25[7]는 검색엔진이 주어진 검색 질의에 대한 문서의 관련성을 추정하기 위해 사용하는 Ranking function이다. 기존의 TF-IDF 알고리즘을 개선한 것으로 확률론적 검색 프레임워크를 바탕으로 한다. BM25는 두 가지의 파라미터로 가중치를 조절할 수 있다. TF(term frequency) 값에 대한 파라미터로 TF 값에 제한을 주어 TF 값이 일정한 범위를 유지하게 하고 문서 길이에 대한 파라미터가 있어 기존의 TF-IDF와 달리 평균 문서의 길이보다 짧은 문서에서 단어가 매칭된 경우 더 높은 가중치를 부여한다. 이 논문에서는 역색인 방법과 BM25를 적용하여 문서를 검색한다.

2.2 BERT

BERT는 공개 이후 자연어 처리의 다양한 분야에서 많은 성능 향상을 이루어낸 사전학습 모델이다. 기존의 정해진 방향대로 학습하던 모델과 다르게, BERT는 양방향으로 학습한다[8]. 일정 비율의 토큰을 가린 뒤, 그 토큰에 해당하는 단어를 예측하며 학습한다. 그리고 Q&A와 같은 작업을 학습하기 위해 다음 문장을 예측하는 학습을 진행한다. 이런 사전학습 과정을 통해 모델이 단어 혹은 문장 간의 관계를 이해하게 된다. 영어뿐 아니라 다양한 언어로 사전 학습된 다국어 버전 BERT도 제공되고 있는데, 다국어 버전 BERT의 경우 한국어 Q&A 데이터셋 KorQuAD[5,6]로 Fine-tuning 해도 좋은 성능을 보인

다. 이 논문에서는 먼저 KorQuAD 1.0을 이용해 보편적인 질의응답 과정을 학습시킨 뒤, 자체적으로 구축한 프로야구 기사에서의 Q&A 데이터 셋으로 강화 학습을 시켜 프로야구 질의응답에 대한 성능을 평가한다.

3. 제안 시스템

3.1 시스템 구조

그림 1은 제안하는 한국 프로야구 Q&A 시스템의 구조를 보인다. 네이버 뉴스의 프로야구 기사[9]를 크롤링하고 JSON 형태로 변환하여 저장한다. 기존 Google-Bert에 질의응답을 진행하기 위하여 KorQuAD 1.0 데이터셋으로 사전 학습 후 프로야구 질문에 특화된 야구 데이터로 강화 학습을 한다. BM25 알고리즘을 이용하여 질문과 기사들 간의 BM25 score를 계산, 관련도 상위 기사를 추출한다. 학습된 BERT 모델에서 질문과 검색된 기사 본문을 입력으로 사용하여 답을 추출한다. 추출한 답은 형태소 분석기[10]를 이용하여 형태소 단위로 나누고 불용어 제거 후 사용자에게 제시한다. 아래에서 각 단계에 대해서 상세히 설명한다.

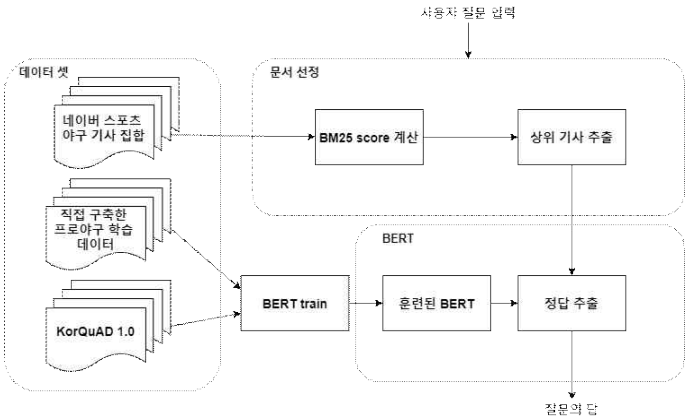

그림 1. 시스템 구조도

3.2 문서 선정

문서 선정은 수집한 네이버 스포츠 뉴스 기사 중 질문과 연관된 기사를 추출하는 모듈이다. 질문과 기사 연관도는 BM25 알고리즘에 기반한 수식 (1)로 계산하며[11], 연관도가 높은 문서를 정답 추출 대상 문서로 선정한다.

$$\text{score}(D,Q)=\sum_{i=1}^{n}\text{IDF}(q_i)\cdot\frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)} \quad (1)$$

수식(1)에서 D는 문서, Q는 질문 q_i는 질문내 토큰에 해당한다. $IDF(q_i)$는 q_i의 역문서 빈도 가중치, $f(q_i,D)$는 문서 D에서 q_i의 빈도를 나타낸다. $avgdl$은 문서의 평균 길이, $|D|$는 문서 길이를 나타내어, 식에서는 문서의 길이가 평균보다 짧을수록 점수가 향상된다. b는 문서의 길이 비율이 검색에 얼마나 영향을 미치는지에 대한 가중치이다. k_1는 용어 주파수 포화 특성을 결정하는 값이다. b와 k_1은 고급 최적화가 이루어지지 않는 경우 $b=0.75$, $k_1\in[1.2,2.0]$의 값으로 책정한다. 질의 용어가 문서에 많이 나타날수록 점수가 높아지는 것이 일반적이다. 하지만, 특정 단어가 많이 나타날 경우 미치는 영향력의 제한을 둔다.

3.3 정답 추출

문서 검색 모듈을 통해 나타나는 결과 중 뉴스 본문을 토대로 정답을 추출하는 것을 정답 추출이라 한다. 질문과 뉴스 본문을 합쳐 최대 512 Token까지로 길이를 제한한다. 먼저 KorQuAD 1.0을 이용해 보편적인 질의응답 과정을 학습시킨 뒤에 자체적으로 구축한 프로야구 기사에서의 Q&A 데이터 셋으로 강화 학습을 수행한 BERT 모델에 질문과 뉴스 본문을 입력으로 사용한다. 모델을 통해 본문 정답의 시작 위치와 끝 위치를 파악하여 정답으로 추출한다.

4. 실험 및 실험 결과

4.1 데이터셋

본 시스템에서는 그림 1과 같이 세 종류의 데이터를 활용한다. KorQuAD 1.0[5,6][1]는 학습집합 60,407개, 검증집합 5,774개의 질의응답 쌍으로 구분하였다. 전체 데이터는 위키피디아 기사에 대한 10,645건의 문단으로 구성되어있다. 문서 검색을 위한 네이버 스포츠 야구 기사 집합은 2019년 1월 1일부터 2020년 9월 15일까지의 11,845개의 기사를 포함한다. 제목, 기사 날짜, 기사 이미지, 본문, 뉴스 링크, 팀으로 구성되어 있다.

KorQuAD는 일반적인 Q&A를 위해 구축되어 야구에 적합한 기계독해 결과를 낼 수 없다. 본 시스템에서는 프로야구 특화 Q&A를 위한 질의응답 데이터를 추가로 구축하였다. 2016년 1월1일부터 2018년 12월 31일까지의 뉴스 기사를 토대로 268개의 데이터셋을 구축하고, 학습 집합 213개, 검증집합 55개의 질의응답 쌍으로 구분하였다. 성능 평가는 직접 구축한 뉴스 기사 검증 셋으로 진행했다. 모든 데이터는 KorQuAD와 동일한 형태로 번호, 제목, 질문, 본문 내용, 본문 내용에 포함된 정답, 정답의 시작 위치로 구성된다. 학습에는 질문, 본문 내용, 정답, 정답 시작위치를 사용한다.

4.2 평가 및 실험 결과

문서 선정의 정확도를 평가하기 위해 연관도 상위 N개의 문서를 선정하여 실제 정답이 존재하는지를 측정했다. 표 1는 수동 평가를 통한 10개 질의 결과를 보인다. 결과에서는 제목보다 본문에서 높은 결과를 얻을 수 있었으며, 대부분 질의에서 5개 문서에서 정답을 찾을 수 있었다. 본 시스템에서 사용자에게 정답을 제공하는 시

표 1. 수동 평가를 통한 10개의 질의-결과 정확도

	N = 5	N = 10	N = 15
기사 제목	44%	35%	21%
기사 본문	72%	71%	57%

1) 제안 시스템은 텍스트 형태 질의에 대한 정답을 추출하는 것이므로 KorQuAD 2.0가 아닌 KorQuAD 1.0을 사용한다.

간을 고려하여 상위 5개 미만의 문서를 사용한다.

본 시스템에서 정답 추출에 두 가지 모델에 대해 성능을 비교했다. 성능 측정은 기계 독해에서 사용되는 지표인 F1 score, EM score(Exact Match score)를 통하여 측정했다. 언어 모델은 Google Bert-Multilingual model을 base model로 활용한다. 파라미터값을 수정해 나가며 최상의 성능을 만들어 내는 모델을 구성하였다. RAdam Optimizer[12], learning rate=1e-5, 활성 함수는 gelu[13], 입력 최대 길이는 512, KorQuAD 1.0의 batch size = 4, epochs = 3, 프로야구 학습 셋의 batch size = 4, epochs = 20으로 진행했다. 측정 결과는 표 2과 같다.

표 2. 모델 성능 측정 결과

	KorQuAD 1.0	KorQuAD 1.0 + Baseball Training set
F1 score	61.13%	76.77%
EM score	51.11%	66.66%

표 2에서 BERT 모델에 KorQuAD 1.0만을 fine-tuning을 진행한 모델보다 직접 구축한 프로야구 학습 데이터셋를 추가로 학습시킨 결과가 더 높은 성능 결과를 보인다. 성능 평가 결과를 비교하였을 때, 야구 용어에 대한 질의와 응답에 대해 더 정확한 정답을 추출한 것이 성능 향상의 근거가 되었다고 판단한다.

그림 2는 구축된 시스템의 화면을 보인다. 질의와 문서 검색 점수가 높은 상위 3개의 기사 순으로 제시한다. 그림의 '삼성 최채흥 데뷔 첫 완봉승 탈삼진 개수는?' 이라는 질문에 실제 답변이 '10개', '10탈삼진' 라는 답변을 제시함을 확인할 수 있다.

그림 2. 프로야구 Q&A 시스템 UI

5. 결론

본 논문에서는 사용자의 질문을 통해 BM25 알고리즘과 기계 독해를 통한 한국 프로야구 질의응답 시스템을 제안했다. 프로야구 데이터를 자체적으로 구축하여 시스템 성능 향상을 비교했다. 기존 KorQuAD 1.0만으로 학습을 진행했을 때 보다 프로야구에 특화된 데이터를 fine-tuning 진행하였을 때 F1 score와 EM(Exact Match)에서 약 15%의 성능 향상을 보였다.

향후 시스템에서는 야구 규칙과 같은 비정형 텍스트 데이터에 대해서도 폭넓은 질의응답 시스템으로 확장할 예정이다. 또한, 추출한 문서의 길이가 입력 길이를 초과하는 경우 문단별로 나누어 정답을 추출하는 방식으로 시스템을 보완할 예정이다.

참고문헌

[1] "엔씨, AI 야구 정보 서비스 '페이지(PAIGE)' 출시", 2018년 07월 24일 수정, 2020년 09월 20일 접속, http://www.ddaily.co.kr/news/article/?no=170959.

[2] 강유환, 안영민, 서영훈, "질의문 키워드의 가중치 부여 방법을 이용한 정답 문서 순위화 시스템", 컴퓨터정보통신연구, Vol.13, No.1, pp.105-110, 2005.

[3] 권순만, 박병준, "단어기반 웹문서 검색을 위한 효과적인 단어 가중치의 계산", 한국정보과학회 학술발표논문집, Vol.31, No.2, pp.169-171, 2004.

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, arXiv:1606.05250, 2016

[5] 임승영, 김명지, 이주열, "KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋", 한국정보과학회 학술발표논문집, pp.539-541, 2018.

[6] KorQuAD, https://korquad.github.io/category/1.0_KOR.html

[7] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford, "Okapi at TREC-3", In Proceedings of the Third Text REtrieval Conference (TREC-3),1994

[8] Jacov Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of NAACL-HLT 2019, pp 4171-4186, 2019.

[9] 네이버 스포츠 뉴스, https://sports.news.naver.com/kbaseball/index.nhn

[10] KoNLPy Twitter, https://konlpy.org/en/v0.4.4/api/konlpy.tag/#module-konlpy.tag._twitter

[11] Okapi BM25, https://en.wikipedia.org/wiki/Okapi_BM25

[12] RAdam, https://github.com/LiyuanLucasLiu/RAdam

[13] Dan Hendrycks, Kevin Gimpel, Gaussian Error Linear Units (GELUs), arXiv:1606.08415v4, 2020