TextRank를 활용한 뉴스 요약 및 MRC 학습을 통한 정답률 향상 방법

배현철

고려대 디지털융합금융학과 3기

itmaster@korea.ac.kr

HYUNCHUL BAE

Korea University, Digital Finance Convergence Engineering

요 약

TextRank를 활용한 뉴스요약 및 Keywords 분석에 대해서 연구하고 실제 뉴스에 대해서 얼마나 잘 요약을 할 수 있는지 확인을 해보았다. 또한 한국어 질의 응답 모델인 KoElectra 모델을 가지고 Korquad의 지문과 질의 답변의 정답률을 최적으로 하는 방법에 대해서 연구를 하였다. 성능 튜닝을 위한 주요 파라미터로는 배치사이즈, 러닝레이트, 에포크 수가 있고 이중 각각의 파라미터가 학습 효과 및 처리 속도에 미치는 영향에 대해서 확인하고, 시간을 효율적으로 할 수 있는 방법을 고민해보았다. 그 외 정답율을 높일 수 있는 방법이 어떤 것들이 있는지 연구하고, 전처리처리와 품사태깅을 활용한 불필요한 단어의 정제가 정답율에 영향을 미치는 방법에 대해서도 연구하였다.

1. 서론

본 연구에서는 두가지 주제를 가지고 진행을 하였다. 첫째는 TextRank를 이용하여 입력된 뉴스요약 과 Keywords 추출에 대한 연구 및 실습이고, 둘째는 KoElectra 모델을 활용한 Korquad 데이터셋에 대한 학습 및 성능향상으로 정답율 개선 과제였다.

넘쳐나는 뉴스와 정보를 일일이 다 읽고 Summary 하고 필요한 정보만 취득할 수 있는 방법이 있다면 무한 경쟁에서 살아 남기가 수월 할 것 같다. 치열한 경쟁에서 가능한 빨리 정보를 액세스 하고 키워드만 출력할 수 있는 시스템이 있다면 시간 절약에 도움이 될 것이다.

이번 과제를 진행하면서 Kaggle 리더보드라는 경쟁시스템으로 연구의 개선 향상을 한 번에 확인을 할 수 있어 성능을 향상하는 부분을 집중했으며, 실제 하고자 하는 여러가지 실험적인 부분은 다소 미흡했다.

다소 어렵고 힘든 부분이 많은 과정임에도 불구하고 새로운 부분을 배움으로서 학습한 내용을 회사의 업무에 사용할 수 있는 부분에 대해서 한번 더 고민하고 필요성 생각 할 수 있는 부분에서는 좋은 기획인 것 같다.

본 소논문에서는 TextRank의 일반적인 동작원리에 대하여는 간단하게 설명하고, Korqard 데이터셋에 대하여 전처리를 수행한 방법과 어떤 방법으로 튜닝을 하면 효율적으로 가능 한지 확인한다.

2. 관련 연구

문서의 요약은 긴 텍스트 문장에서 의미 있는 데이터를 추출 또는 요약된 문장을 생성하는 것을 말한다. 입력된 텍스트를 분석하고 정제하고 의미 있는 문장으로 만들어 읽는 사람들이 이해할 수 있어야 한다. 여기서 말하는 문서요약의 방법은 추출적 문서요약을 말한다.

추출적 문서요약 방법은 원문에서 중요한 핵심 문장 또는 단어구를 몇 개 뽑아서 이들로 구성된 요약이다. 대표적인 알고리즘으로 머신러닝 알고리즘인 텍스트랭크(TextRank)이고 추출형 문서요약의 장점으로 원 문서에 있는 문장을 선택하여 요약문을 생성하므로 구현이 비교적 쉽고 문법적으로 완성이 되어있는 문장들을 추출하므로 문법적으로 오류가 날 가능성이 적다.

기계독해(MRC)는 인공지능(AI)알고리즘이 스스로 문제를 분석하고 질문에 최적화된 답안을 찾아내는 기술을 이야기 한다. 대표적인 MRC인 QA(Question Answer)시스템처럼 인간이 텍스트를 읽으면 AI가 답변을 추론하여 답하는 것이 기존의 조건별 응답시스템 (DB 형, 키워드기반 응답시스템) 보다 유연한 답

을 주고있다.

3. 데이터 특성

이번 MRC 과제에서 사용한 데이터 셋은 Korquad는 Korean Question Answeering Dataset으로 한국어 Machine Reading Comprehension를 위한 데이터셋이다. 전체 데이터는 1560개 WiKipedia 기사에 대해 10645개의 질의 응답 쌍으로, Training set 60407개, Dev Set 5774개의 질의 응답 쌍으로 구분 되어있다. TextRank는 특정 뉴스에 대한 입력 값에 대한 처리이기 때문에 특별한 데이터 셋은 없고, 특정 뉴스 URL로 대처해서 테스트를 진행하였다.

4. 개발 방법론

4-1 데이터전처리

Tokenizer 만 사용하여 진행을 하면 단어들의 끝에 불필요한 조사 및 불용어들이 붙어 있어 EM점수 및 F1 점수를 낮게 받을 수 있는 문제가 발생한다. 특히 한국어는 어미와 조사 등이 발달되어 있기 때문에형태소에 따라 단어를 분리하고 불필요한 단어를 제거 함으로서 학습된 데이터의 정답율을 높일 수 있다.
여러 종류의 한국어 형태소 분석기가 있는데, 대표적인 형태소 분석기를 활용 해보고 각각의 장단점에 따른 결과들을 확인 해본다.

문장 : 사랑하고싶게하는가슴
태깅 : '사랑/NNG', '하/XSV', '고/EC', '싶/VX', '게/EC', '하/VV', '는/ETM', '가슴속/NNG'

위의 예시는 **Mecab형태소** 분석기를 활용하여 품사태깅을 한 예시이다.

전처리에서 한글,영문,숫자만 허용했고, 특수문자와 한자는 ㄱ-ㅎ가-힣 으로 지정해 ㄱ-힣 내의 한자를 제외했다.

4-2 학습

실습으로 제공되어진 학습모델을 가지고 진행을 했으며 필요한 주요 파라미터를 변경해가면서 모델의 성능에 대한 향상여부를 연구하였다.

4-3 평가방법

Korquad 에서는 모델을 평가하기 위한 평가 스크립트와 입력 샘플 예측 파일을 제공하고 있고, Dev set에 대해 만족하는 모델을 평가하기 위해서 리더보드에 제출하면 Test Set으로 평가 한 Exact Match(EM) 및 F1 Score로 점수를 받는다.

5. 모델

먼저 TextRank 에 대한 내용을 간략하게 설명한다.
TextRank 는 아래 3가지로 진행이 된다.

전처리 과정 : 문서를 문장단위로 분리하고 명사추출
TF-IDF : 각단어들의 가중치를 계산
TextRank 적용 : 중요도가 높은 순으로 정렬

TextRank는 단위별로 제공된 모델을 가지고 연구를 하여 실제 뉴스에 대한 정확도 및 실효성에 대해서 확인을 해보았다. 요약의 정확성에 대한 부분과 Keywords들의 연관성에 대해서도 확인해보았다.

KoElectra는 BERT 이후에 등장한 언어모델로서, BERT가 가진 학습 데이터 사용의 비효율성을 극복하기 위해 탄생한 모델이다. 기존 모델인 BERT는 학습 과정에서 전체 입력 토큰 중 [MASK]로 가려진 15%의 토큰들만 학습에 사용하기 때문에 데이터 효율성이 떨어지게 된다. ELECTRA는 이를 극복하기 위해 [MASK]로 가려지지 않는 나머지 85% 토큰에 대해서도 학습을 진행하므로 BERT대비 초기 학습 속도와 성능 면에서 우수하다고 증명되었다. 이러한 ELECTRA 모델 아키텍처를 한글로 학습한 모델이다. 이 모델은 무겁지만, 더 좋은 성능을 자랑한다.[참조 1]

6. 실험환경

구글의 Colab 환경을 사용하였다. GPU의 유무에 따라서 학습하는 속도에 큰 차이를 보였고, 유료결제를 하면 Colab PRO을 사용할 수 있고 더 빠르고 편리한 환경에서 사용이 가능하였다.
Colab PRO = 고용량 RAM(25.51 G) , RunTime 유지시간 (24시간)

데이터 셋은 Korquad는 Korean Question Answeering Dataset으로 한국어 Machine Reading Comprehension를 위한 데이터셋이다. 전체 데이터는 1560개 WiKipedia 기사에 대해 10645개의 질의 응답 쌍으로, Training set 60407개, Dev Set 5774개의 질의 응답 쌍으로 구분 되어있다.

7. 하이퍼 파라미터

이번 과제에서 하이퍼 파라미터로 Batch Size, Epochs, Learning Rate 세가지가 성능에 큰 영향 요소이다.
또한 작업에서 중점으로 확인했던 부분은 전처리, 조사제거에 대한 형태소분석 부분으로 해당 모델도 정답율에 영향을 미쳤다
초기 epoch = 10 , Batch size = 32 , Learning Rate = 5e-5 로 진행
Konlpy의 Hannanum, Mecab, Kkma, Komoran 등이 있다.
각 형태소 분석기를 사용하면서 장단점을 확인해 보았다.

8. 정답율을 높이기 위한 튜닝

실제 모델에 영향을 줄 수 있는 Factor 로 형태소 분석기를 활용한 전처리 부분과 하이퍼 파라미터로 생각하고 연구를 계획하였다.
형태소 분석기 별로 속도 및 결과가 확실히 차이가 발생하였고 결과에 대한 확인을 확실히 확인이 되었지만, 하이퍼 파라미터의 경우 다양한 연구를 진행하지 못하였다.
Batch size / Epoch 처럼 수행 속도에 영향을 주는 경우가 있고, Learnning rate 처럼 수행 속도에는 영향을 안주는 경우가 있었다.
파라미터들을 변경해 나가는 방식으로 테스트를 진행해 보았고 별도로 형태소분석기도 바꿔가면서 진행을 해보았다.

8-1 튜닝 결과 형태소분석기

형태소분석기를 여러가지로 바꿔가면서 정답율에 대한 변화를 테스트를 해봤고 정답률과 속도에서 차이가 다르게 나왔다. 실제 가장 좋은 결과를 보인 분석기는 Hannanum 로 가장 좋은 결과를 보였다. 단 속도는 다른 분석기에 비해서 차이가 많이 났다. 속도측면에서는 Mecab 분석기가 가장 빠르게 반응을 하였다. Mecab는 속도는 빠르게 나왔지만 정답율은 Hannanum 분석기보다는 다소 떨어졌다.

8-2 튜닝 결과 MRC

learning rate - Epoch 의 관계에 대해서 테스트를 진행 하였고, epoch의 수와 Rate의 변화를 통해서 가장 좋은 성능을 보이는 파라미터 셋을 찾아야 했다. 최초 계획대로 모든 케이스를 다 진행할 수는 없어서 다소 미흡한 점이 있었지만 두 파라미터의 상관관계에 대해서 알 수 있는 연구를 했다.

9. 결론

TextRank 모델을 이용하여 문서 혹은 뉴스 요약에 대해서 구현을 해보고 TextRank의 결과에 대해서 얼마나 잘 요약하는지 확인해보았다.

KoElectra 모델을 활용한 MRC에서는 성능튜닝을 위한 주요 파라미터로 Batch size,learning rate, epoch 가 있고 이런 하이퍼 파라미터의 조정으로 튜닝의 최적값을 찾을 수 있다. 또한 데이터의 전처리도 중요한 부분이다.

이번 과제를 진행하면서 MRC에 대해서 이해하는 과정이 되었으며, 모델의 이해와 하이퍼 파라미터의 관계에 따라 성능에 차이에 대해서 연구할 수 있어서 좋은 경험이 된 것 같다
과제의 최적 파라미터를 찾기 위해서 수십번의 과정을 반복하여 작업을 수행하였지만 성능이 확 좋아지는 최적의 값을 찾지는 못하였다. 단 형태소 분석기라는 부분에 대해서는 정확한 학습을 할 수 있는 기회가 된 것 같다. 형태소 분석기에 따라서 성능이 차이가 날 수 있고, 속도에도 영향을 미치는지 확인이 되었다. 추후 이런 프로젝트를 하게 된다면 이런 수작업이 아니라 최적의 값을 찾는 학습모델을 만들어 보는 것도 좋은 과제가 될 것 같다. 이번 과정을 진행하면서 많은 부족한 부분이 존재하는 것을 인지했고, 향후 나은 방법을 위해서 모델링 기법과 다양한 접근법으로 처리법 등을 익혀야 할 것 같다.
특히 이번 과제를 통해서 의미 있었던 것은 비록 기초적인 지식을 바탕으로 실습과제를 기준으로 원우분들과 경쟁을 하면서 어떻게 성능을 튜닝을 하면 좋은 결과가 나오는지? 튜닝단계는 어떻게 계획을 해야 하는지? 파라미터에 대한 더 깊은 배움이 필요하다고 생각을 하였다.

종합적으로 성능 튜닝을 위한 주요 파라미터로는 Batch size, learning rate, epoch 가 있고 이런 하이퍼 파라미터의 조정이 성능을 높여 줄 수 있다. 또한 형태소 분석을 통한 전처리 처리가 학습시간에 많은 영향을 주므로 적절히 설정해야 한다.

참고문헌

[1] https://www.samsungsds.com/kr/insights/TechToolkit_2021_KoreALBERT.html
[2] https://joolib.tistory.com/21
[3] https://bab2min.tistory.com/552
[4] https://koreascience.or.kr/article/JAKO201835146902030.pdf