

도메인 지식에 특화된 한국어 대규모 언어모형 기반 챗봇



정보컴퓨터공학과 201624597 최상준

정보컴퓨터공학과 20184448 김정민

불어불문학과 201902163 김채원

지도교수 권혁철

목 차

1. 서론.....	1
1.1. 연구 배경.....	1
1.2. 기존 KoAlpaca 모델의 문제점.....	1
1.3. 연구 목표.....	2
2. 당초 연구 계획.....	2
2.1. 지식인 데이터셋.....	2
2.1.1. 지식인 크롤링.....	2
2.1.2. 지식인 데이터 필터링.....	2
2.2. 키프리스(Kipris) 데이터셋.....	3
2.3. 배포가능한 프롬프트 구현.....	3
3. 계획 수정 사항.....	3
3.1. 지식인 필터링 방식 수정.....	3
3.1.1. 기존 방식의 문제점.....	3
3.1.2. 개선 방식.....	6
3.2. 키프리스 데이터 크롤링 방식 수정.....	6
3.2.1. 기존 방식의 문제점.....	6
3.2.2. 개선 방식.....	7
4. 연구 결과 분석 및 평가.....	8
5. 결론 및 향후 연구 방향.....	8
5.1. 결론.....	8
5.2. 향후 연구 방향.....	9

6. 참고 문헌.....	9
---------------	---

1. 서론

1.1. 연구 배경

2022년 ChatGPT가 큰 인기를 얻음에 따라 대화형 인공지능 서비스에 대한 관심이 높아지고 있다. 국내에서는 Stanford Alpaca 모델을 차용한, 한국어를 이해하는 KoAlpaca가 공개되어 주목을 받기도 했다.

1.2. 기존 KoAlpaca 모델의 문제점

하지만 이를 실제로 사용해보았을 때 몇 가지 문제점이 있었는데 이는 문장 구성이 어색하거나, 문맥을 잘 파악하지 못한다는 부분이다. 전문성이 요구되는 영역에서 답변의 질이 상당히 떨어진다는 것이 가장 취약점이다. 해당 서비스는 네이버 지식IN을 크롤링하여 제작되었다. 해당 웹사이트의 특성상 의학, 법학, 어문학 등의 깊은 학문적 지식을 얻는 것은 한계가 있다. 다음의 사진과 같이 특허 부문에 관한 질문을 던지자 실제로 존재하지 않는 특허를 답변으로 제시했다.



<사진 1> KoAlpaca의 문제점

1.3. 연구 목표

이러한 문제점을 해결하기 위해 특정 부문에 특화된 한국어를 사용하는 챗봇을 구현하고자 했다. 그 중 특화 도메인은 특허 부문으로 선정했다. 한국어를 사용하는 특허 정보를 담은 챗봇을 통해 특허를 출원하고자 하는 이, 특허 검색이 필요한 이들이 현재보다 편리하게 원하는 정보를 얻을 수 있는 서비스를 구현하는 것을 목표로 했다.

2. 당초 연구 계획

2.1. 지식인 데이터셋

2.1.1. 지식인 크롤링

지식인에서 제공하는 여러 서비스 중 이번 연구의 크롤링에 사용할 데이터는 '베스트 문답'과 '분야별 문답' 두 가지이다. 베스트 문답은 지식인에서 선정한 놀랍고 재밌는 질문 및 답변을 말하는데 여기에는 다양한 주제의 질문이 있다. 특히 지식인에 의해서 한 번 걸러진 게시글이므로 데이터의 질이 높을 것이라 판단하여 이를 선정했다. 실제로 여러가지 주제에서 다양한 질문 및 답변 데이터를 얻을 수 있었다. 또한 비속어, 광고성 게시글, 무성의한 답변 등이 없기 때문에 이후의 데이터 학습에서 비교적으로 매끄러운 결과를 얻을 수 있었다.

분야별 문답은 각종 분야 별로 답변을 많이 게시한 이용자를 모아서 볼 수 있는 탭이다. 지식인에서 활발히 활동하는 사람들이기 때문에 답변의 질이 높을 것이라 판단했다. 또 다양한 분야에서 제각기 다른 말투를 구사하는 사람들의 텍스트를 얻을 수 있어 이후 학습에서 좋은 결과를 기대할 수 있을 것으로 보아 선정했다.

베스트 문답에서는 2.3만 개, 분야별 문답에서는 80만개의 데이터를 크롤링 했다.

2.1.2. 지식인 데이터 필터링

크롤링 계획 수립 시에 질문과 답변의 질을 고려하여 데이터를 선정하였지만, 검증되지 않은 사용자들에 의해 작성된 텍스트이기 때문에 데이터 필터링이 필요했다. 이때 데이터 필터링에는 파이썬 Re 라이브러리를 사용하기로 계획했지만 이후에 몇 가지 문제점이 발견되어 수정되었는데 문제점과 해결 방안은 아래의 수정사항 항목에서 후술하겠다. 먼저 당초 계획을 간략히 설명하고자 한다. 해당 라이브러리는 대량의 텍스트 사이에

서 특정 부분을 검색 및 추출하기 위해 주로 사용된다. 이번 연구에 적합한 데이터를 학습할 수 있도록 필터링한 기준은 다음과 같다.

A. 특정 키워드가 들어간 질문 혹은 답변 전체 제외

예시: http, HTML (개인 홍보), 남남 (지식인에서만 사용되는 용어), 닥치 (비속어)

B. 반복되는 특정 문구만 제거

예시: '안녕하세요. 하이닥-네이버 지식in 상담의 ○ ○ ○ 입니다.

C. 답변의 단어 수 2개 미만 시 답변 전체 제외

2.2. 키프리스(Kipris) 데이터셋

키프리스는 특허정보검색서비스이다. 해당 사이트에서 제공하는 엑셀 파일을 저장하여 크롤링한다. 크롤링할 주요 내용으로는, 특허의 요약 섹션 및 저자 등이 있다. 이미 질이 검증된 데이터들이므로 따로 필터링은 필요하지 않다. 하지만 학습에 사용할 수 있도록 json 파일로 변환하는 과정을 거쳐야 한다.

연구를 진행하는 동안 “한 문제점을 발견해 크롤링 방식을 변경했다. 이는 후술한다.

2.3. 배포가능한 프롬프트 구현

이거 계획에 넣어줘야할까요?

3. 계획 수정 사항

3.1. 지식인 필터링 방식 수정

3.1.1. 기존 방식의 문제점

기존 계획과 같은 방식으로 데이터 필터링 후 학습시켰을 때 (이때 데이터들은 Polyglot-ko-1.3b로 1에폭 학습했다.) 드러난 문제점을 크게 나누면 다음의 세 가지 유형으로 볼 수 있다.

a. 답변이 긴 경향

지식인 데이터를 요약하지 않은 채 기본적인 키워드 필터링 및 부적절한 데이터 전체 제외했다. 요약하지 않은 데이터들을 학습했기에 학습 결과의 답변도 길이가 긴 경향을 보였다.

b. '질문', '답변', '내용' 등의 단어 사용

파인튜닝의 영향을 받은 것으로 파악했다. 이는 학습시킨 데이터를 충분히 학습하여 출력된 결과로 보였다.

c. 잘못된 / 편향된 답변

다양한 답변자가 존재하기 때문에 허위 사실 혹은 편향된 데이터가 많이 존재한 것을 원인으로 보았다.

아래는 예시이다.

```
##질문:낙타가 뭐야?
##답변:낙타는 사막에서 살면서 물이 없는 사막에 사는 동물인데요.
낙타는 크게 3가지 종류가 있습니다.
털이 없는 사막여우
털이 있는 사막 여우
사막에서 사는 사막 여우는 사막에 사는데, 사막 여우가 사는 곳은 사막의 모래에는 풀이 자라지 않습니다.
그래서 사막에 사는 사막여우는 사막의 모래에 풀을 먹기 위해서 낙타를 이용합니다.
낙타가 사는 곳을 사막이라고 합니다.
털이 있거나 없거나 사막에서 살기 때문에 사막 여우도 사막에서 사는데, 사막에 사는 낙타는 사막의 모래
에서 풀을 먹기 위해 낙타를 사용하는 것입니다. 사막에 사는 여우는 사막에 풀을 먹고 사는 낙타를 이용해
서 사막에서도 살 수 있습니다.

도움이 되셨다면 채택 후 내용의 50%를 돌려받으세요.

채택 받은 답변은 질문자의 답변으로 질문자의 명예나 불이익을 위에 알려드리지 않으며,
채택을 해주셔도 답변 한 분에게는 해피 콩 1
```

<사진2> a, b에 해당하는 경우

##질문:전라도 사람은 어때?

##답변:서울사람들도 전라도사람이라면 싫어해요.

전라도 사람들은 같은 지역에 사는 사람들하고만 친하죠.

전라도사람들은 겉으로는 전라도 사람인척하면서 속으로는 서울사람이랑 놀아요.

전라도를 욕하는것은 전라도 사람들도 마찬가지예요.

전라도는 겉으로는 경상도를 욕하면서,

실제로는 경상도 사람들을 굉장히 싫어하죠.

전란때, 전라도 사람들에게 도움을 많이 받았기때문에,

전라도 사람들이 더 마음이 편할거예요.

전라는 겉으로는 경상도 욕하면서, 실제로는 경상도랑 친하죠.

전라도 사람이 서울에 가면 서울사람들을 싫어하고,

전라도사람들이 서울에 가면 전라도 사람들을 싫어합니다.

전라도가 서울에 가면 서울 사람들이 싫어해요.

전라는 겉모습만 경상도를 욕하는척하면서,

실제적으로는 경상도 사람들 엄청 싫어합니다.

<|endoftext|>

<사진3> a, c에 해당하는 경우

이에 대해 산학협력 멘토의 의견서는 다음과 같았다.

1. 지도개요

팀 명	응답하라 챗봇
과 제 명	도메인 지식에 특화된 한국어 대규모 언어모형 기반 챗봇
협력기관	(주)나라인포테크

2. 세부 지도 내용

데이터 전처리와 1 차 학습을 통해 보완점을 잘 파악함.
추가 전처리를 통해 현재 발견된 문제점들을 보완할 수 있을 것으로 파악됨.

아래는 이후 진행 상황에서 고려해 볼 만한 문제점 해결 방안입니다.
답변이 긴 경향: 학습 데이터의 길이를 제한하는 방식이 도움이 될 수 있음. Ex) KoBertSum 을 사용해 핵심 문장만으로 구성된 축약된 답변을 생성하여 학습 진행.

'질문' '답변' '내공'등의 단어 사용 또한 앞에서 진행한 전처리와 유사하게 해당 문장의 중요도를 고려한 문장 필터링을 거쳐 제거하면 답변의 완성도가 올라갈 것으로 예상됨.

잘못/편향된 답변 -> 학습 데이터에서 편향된 답변을 파악하고 사전에 제거 (ex> 성별, 지역, 인종 등에 대한 키워드 위주로 인식하여 필터링), K-StereoSet 과 같은 편향성 인식 모델을 활용.

최종 결과에서는 답변의 정확도를 BLEU, ROUGE 등의 지표를 통해 수치적으로도 확인할 수 있으면 완성도가 높아질 것 같습니다.

3.1.2. 개선 방식

GPT 이용

상술 필요

베스트 문답과 분야별 문답의 데이터셋 개수를 합하면 약 82.3만 개이다. 필터링 이후에는 약 10만 개의 데이터셋이 남았고 이것들로 최종적 학습시켰다.

3.2. 키프리스 데이터 크롤링 방식 수정

3.2.1. 기존 방식의 문제점

어떤 문제점이었나요??

3.2.2. 개선 방식

엑셀 파일 대신 PDF파일을 이용하기로 했다. 먼저 카테고리 별 PDF파일 다운로드 한다.

Smart Node에서의 All-IP 기반의 Autonomous 모바일 커뮤니케이션 방법 및 시스템 < 통합행정정보 - 특허 실용신안 정보 - Mo... >

Smart Node에서의 All-IP 기반의 Autonomous 모바일 커뮤니케이션 방법 및 시스템
System and method for All-IP based Autonomous Mobile Communications on Smart Node

상세정보 공개전문 공고전문 등록사항 통합행정정보

통합행정정보

본 '원본보기 서비스'는 참고용이므로, 일부 오류 및 누락이 발생할 수 있습니다.
정확한 서류를 확인하시려면 해당 웹사이트에서 조회하시기 바랍니다. (특허로 바로가기: <http://www.patent.go.kr>)
해당 서비스는 정검으로 인해 매주 월요일 00:00 ~ 02:00까지 이용이 중단됩니다.

번호	서류명	접수/발송일자	처리상태	접수/발송번호
1	【특허출원】특허출원 ([Patent Application] Patent Application)	2008.11.24	수리 (Accepted)	112008080867450
2	출원인정보변경(경정)신고서 (Notification of change of applicant's information)	2003.08.04	수리 (Accepted)	412009515089936
3	선행기술조사요청서 (Request for Prior Art Search)	2010.12.13	수리 (Accepted)	919999999999999
4	선행기술조사보고서 (Report of Prior Art Search)	2011.01.17	수리 (Accepted)	912011000331517
5	의견제출통지서 원본보기 (Notification of reason for refusal)	2011.01.19	발송처리완료 (Completion of Transmission)	952011003481734

키프리스에 접속하면 위와 같이 특허 정보의 원본을 PDF파일로 다운받을 수 있다.



특허출원서

【참조번호】	0005
【출원구분】	특허출원
【출원인】	
【명칭】	한국전자통신연구원
【특허고객번호】	3-1998-007763-8
【대리인】	
【명칭】	특허법인 무한
【대리인번호】	9-2007-100061-4
【지정된변리사】	구기완
【포괄위임등록번호】	2007-052305-1
【발명(고안)의 국문명칭】	Smart Node에서의 All_IP 기반의 Autonomous 모바일 커뮤니케이션 방법 및 시스템
【발명(고안)의 영문명칭】	System and method for All_IP based Autonomous

해당 PDF파일들은 위와 같은 형식으로 구성 돼 있다. 학습에 사용하기 위해 해당 파일을 텍스트로 변환하여 필요한 부분만 추출한 후 json파일로 변환하여 학습에 사용할 수 있도록 변환하였다.

위와 같은 방식으로 약 10만 개의 특허 정보 데이터를 얻었다.

4. 연구 결과 분석 및 평가

결과 나오면 작성해야 할 것 같습니다.

5. 결론 및 향후 연구 방향

5.1. 결론

결과 나오면 작성해야 할 것 같습니다.

5.2. 향후 연구 방향

1. 현재 진행한 연구는 데이터들을 지식인과 키프리스에서 데이터를 수집했다. 이러한 데이터들은 법적 제약사항이 존재할 수 있으므로 상용화를 위해서는 이 부분을 해결해야 할 것으로 보인다.
2. 이번에 진행한 연구에서는 지식인 데이터로 한국어를 학습시켰다. 조금 더 자연스러운 한국어를 구사하게 만들기 위해서는 지식인 외의 서적이거나 다른 종류의 데이터들을 학습에 사용하면 보다 나은 답변을 얻을 수 있을 것으로 보인다.

6. 참고 문헌

<Web site>

[1] NAVER 지식 IN, <https://kin.naver.com>

[2] KIPRIS 특허정보검색서비스 특허무료검색서비스, www.kipris.or.kr/khome/main.jsp