

# 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. Polyglot-Ko 1.3B 언어모델.....	3
2.4. 배포 가능한 챗봇 구현.....	3
3. 계획 수정 사항.....	4
3.1. 지식인 필터링 방식 수정.....	4
3.1.1. 기존 방식의 문제점.....	4
3.1.2. 개선 방식.....	6
3.2. 키프리스 크롤링 자료 변경.....	7
3.2.1. 기존 자료의 문제점.....	7
3.2.2. 개선 방식.....	7
4. 연구 결과 분석 및 평가.....	9
4.1. 기본 모델.....	9
4.1.1. 일반 질문.....	9

---

4.1.2. 특허 질문.....	10
4.2. 기본모델 + 파인튜닝 .....	11
4.2.1. 일반 질문.....	11
4.2.2. 특허 질문.....	11
4.3. 사전학습 .....	12
4.3.1. 일반 질문.....	13
4.3.2. 특허 질문.....	13
4.4. 사전학습 + 파인튜닝 .....	14
4.4.1. 일반 질문.....	14
4.4.2. 특허 질문.....	15
4.5. 연구 결과 분석.....	17
5. 결론 및 향후 연구 방향 .....	17
5.1. 결론.....	17
5.2. 향후 연구 방향.....	18
6. 참고 문헌.....	18

---

## 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.

이것은 진행 전의 계획으로, 연구를 수행해감에 따라 기존에 사용하려고 했던 엑셀 파일에는 상세한 정보가 없었기 때문에 충분하지 않다고 판단했고 수집하는 자료를 PDF로 변경했다. 이는 후술한다.

## 2.3. Polyglot-Ko 1.3B 언어모델

863GB의 한국어 데이터로 학습된 언어 모델이다. 학습 시 사용된 데이터들의 출처는 블로그 포스트, 뉴스, 특허 정보, 지식인 등의 다양한 소스가 있다. 자기회귀적인 언어 모델로 훈련되었으며, 다음 토큰을 예측하는데 대한 가능도를 최대화하기 위해 교차 엔트로피 손실을 사용하였다.

데이터 필터링 이후 해당 모델을 이용해 학습을 진행했다.

## 2.4. 배포 가능한 챗봇 구현

Gradio 라이브러리로 학습된 모델을 활용하여 챗봇 형태를 구현하고 이를 통해 사용자

---

들이 질문-답변을 진행할 수 있도록 하고자 한다.

### 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

Fine tuning의 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 모래에서 풀을 먹기 위해 낙타를 사용하는 것입니다. 사막에 사는 여우는 사막에 풀을 먹고 사는 낙타를 이용해서 사막에서도 살 수 있습니다.

도움이 되셨다면 채택 후 내공의 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. 지도 개요

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.

세부 지도 내용을 한 번에 해결할 수 있는 방법으로 채택한 것을 다음 3.1.2. 개선 방식에서 서술하고자 한다.

#### 3.1.2. 개선 방식

베스트 문답과 분야별 문답의 데이터셋 개수를 합하면 약 82.3만개이다. 이를 기반으로 텍스트 랭크와 gpt3.5를 이용해 동시에 전처리를 진행했다. 텍스트랭크를 이용한 결과, 전처리 이후 37만 개의 데이터셋이 남았다고 gpt를 이용한 결과 10만 개의 데이터셋이 남았다.

학습 결과 텍스트랭크는 핵심적인 문장만 추출하여 답변의 문장 수를 줄이고 re라이브러리로 처리하지 못한 키워드를 제거하는 데에는 성공했으나 답변의 질이 학습에 적합하지 않았다. Gpt는 문장이 평균적으로 길어지긴 했으나 답변의 질이 높아졌다. 그렇기에

최종적으로는 gpt를 이용한 전처리 방식을 채용했다. Stanford Alpaca와 유사한 방식으로 학습 모델을 완성했다고 할 수 있다.

## 3.2. 키프리스 크롤링 자료 변경

### 3.2.1. 기존 자료의 문제점

이는 기술적인 문제는 아니었다. 다만 당초에 크롤링하고자 했던 엑셀파일에는 학습에 사용할 만한 정보가 충분하지 않았기에 크롤링하는 대상을 엑셀파일이 아닌 pdf파일로 변경했다.

### 3.2.2. 개선 방식

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



The screenshot shows the KIPRIS website interface. On the left, there is a sidebar with search filters and a list of patent numbers. The main area displays a table of patent records. The first record is highlighted with a red box around the '원본보기' (View Original) link.

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	선행기술조사뢰서 (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.



Korean Intellectual Property Office

## 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

기본 모델, 기본 모델 + 파인튜닝, 사전학습, 사전학습 + 파인튜닝 네 가지 모델로 연구해보았다.

1. 기본 모델: Polyglot-Ko-1.3b 1에폭 학습
2. 기본 모델 + 파인튜닝: Polyglot-Ko-1.3b 1에폭 학습 + 네이버 데이터 학습
3. 사전 학습: 키프리스 데이터 학습
4. 사전 학습 + 파인튜닝: 키프리스 데이터 학습 + 네이버 데이터 학습

각 모델에 일상적인 질문과 특허에 관한 질문을 했고 결과물은 다음과 같다.

### 4.1. 기본 모델

한글 생성은 잘하지만 문장 구성이 잘 되지 않는다.

#### 4.1.1. 일반 질문

- 1) 건강하게 살 수 있는 방법 질문

Q: 건강하게 살 수 있는 법 세 가지는?

A: 먹는 것입니다. 먹어야 건강하게 오래 삽니다. 음식은 오래 씹고 씹어야 소화됩니다. 그리고 우리가 많이 섭취하는 음식은 몸에 해롭습니다. 음식은 몸에 좋은 것을 골고루 섭취하세요. 그리고 내가 먹는 것은 내가 직접 만드세요. 그리고 남이 해주는 음식으로는 좋은 것이 없습니다. 직접 만들어 먹는 음식은 몸을 건강하게 합니다. 예를 들어 현미를 백미로, 콩, 채소, 생선, 고기 등을 골고루 먹고, 내가 먹고 싶은 것을 직접 만들어 먹으면 위와 장을 건강하게 하여 면역이 강화되고, 암을 예방, 억제할 수 있습니다. 그리고 마음을 행복하게 하는 것을 먹으면 마음이 행복해지고, 그 행복은 자신이 만들어 먹는 것입니다. 예를 들어 좋은 음악을 들으면 마음이 행복해지는데, 그것은 내가 만들어 먹는 음악이라 할 수 있습니다. (후략)

---

## 2) 낙타에 관한 질문

Q: 낙타가 뭐야?

A: 낙타도 말처럼 순전히 말이 아니고 낙타가 되는 훈련을 한 다음에 낙타인 것입니다.

Q: 소, 돼지, 닭, 개, 염소, 오리도?

A: 낙타라고 하면 우리가 흔히 알고 있는, 낙타가 아니고 낙타 같은 말인 낙타가 아니에요. 낙타 같은 염소나 소, 돼지, 닭 같은 동물들을 말해요.

Q: 그럼 독수리는? 독수리를 낙타라고 하나?

A: 아니에요. 독수리에요

### 4.1.2. 특허 질문

#### 1) 자동차 특허

Q: 자동차에 관한 특허에는 뭐가 있어?

A: 그야 당연히 '오토'죠..^^)\* '오토'의 뜻: 차를 '길들이다'의 뜻.즉 오토차량은 '길들이는 차'라는 말이 되겠고, '길들이다(수동)'의 뜻은'길들이는 과정에서 조작하기 간편한 차량을 말한다. 예를 들어 '수동변속기 차량은 운전자가 차를 처음 운전할때의 상태로 돌아가게 도와주는 기능이 있다'라고 생각하면 된다.물론, 오토차량도 그 기능을 할 수 있지만, 자동화된 기능은 '오토'가 아닌 '수동'이다.\*'오토'의 뜻(예:자동변속기): 자동차의 운전자가 차를 길들이는 과정에서의 조작하기 간편함을 의미함.(즉 수동이 아닌 자동)\* (후략)

#### 2) 노즐 특허

Q: 노즐에 관한 특허 추천

A: AURA WAVE2. AURA TRADA3. AURA LINK4. AURA FLASH5. AURA RADIANCE 6. AURA HEART7. AURA MOVE8. AURA KILL9. AURA SHUFFLE10. AURA PANTS11. AURA SPACE12. AURA DIMENSIONAL13. AURA STICK14. (후략)

---

## 4.2. 기본모델 + 파인튜닝

일반 질문에 대한 답변은 가능하지만 특허에 대해서는 잘 알지 못한다.

### 4.2.1. 일반 질문

#### 1) 건강하게 살 수 있는 법

Q: 건강하게 살 수 있는 방법 세 가지는?

A: 건강하게 살아가기 위해서는 세 가지 방법을 고려해야 합니다. 건강한 신체를 유지하고 건강하게 살기 위해서는 운동과 올바른 영양정보를 섭취하는 것이 필수적입니다. 또한, 올바른 신체활동을 통해 건강을 유지해야 합니다. 활동적인 태도로 운동하는 것도 건강에 도움이 됩니다. 그리고 건강한 식습관을 유지하며 좋은 음식을 섭취하는 것도 중요합니다. 이는 건강한 삶을 유지하기 위한 중요한 요소입니다. 건강한 삶을 위해 올바른 영양정보, 운동, 올바른 식습관이 필요합니다.

#### 2) 낙타에 관한 질문

Q: 낙타가 뭐야?

A: :낙타는 사막에서 많이 사는 동물로, 사막 지역에 살며 등에 낙타의 긴 갈기와 같은 털이 있어서 사막에서 이동할 때 많이 이용됩니다. 낙타는 사막을 이동하며 생존하기 위해 에너지를 절약하고, 물을 절약하는 등의 역할을 합니다. 낙타에는 동물의 특징과 생태가 매우 다양하며, 인간과 공생하는 동물 중 하나입니다. 낙타는 주로 사막에서 살지만, 초원에도 살고 사막의 적도 근처의 작은 산의 평원에도 사는 등 다양한 지역에서 서식합니다. (후략)

### 4.2.2. 특허 질문

#### 1) 노즐 특허 추천

Q: 노즐에 관한 특허 추천

---

A: 최근에노즐에 관한 다양한 연구가 이루어지고 있습니다.<|endoftext|>

## 2) 자동차 관련 특허

Q: 자동차에 관한 특허를 내려고 하는데 어떻게 이미 있지?

A: 이미 있습니다. 이미 있는 특허를 활용하는 겁니다. 자동차에 대한 특허가 있으면 자동차의 모양, 바퀴의 모양, 바퀴굴림, 엔진, 변속기, 기어, 시동거는 방법, 배터리, 기타 등등의 것들을 적용할 수 있습니다. 그리고, 이를 자동차에 응용하면 자동차를 개발하는데 소요되는 시간과 비용을 줄일 수 있고, 자동차의 기능과 성능도 향상시킬 수 있습니다. 자동차에 대한 지식이 없는 사람이라면 아무도 자동차에 관한 지식이 없기 때문에 자동차의 모양, 엔진, 바퀴의 모양과 크기, 변속기의 모양, 기어의 모양, 제동장치의 모양, 시트, 시트에 앉는 방법 등등을 응용할 수 있는 것들에 대해 개발하는데 많은 시간과 비용을 투입해야 합니다. (후략)

## 3) 드론 특허

Q: 드론에 관한 특허에는 뭐가 있어?

A: 드론은 작아서 공간적인 제약이 없습니다. 그래서 드론을 활용하는 방안의 폭이 넓습니다. 저는 드론을 활용해서 많은 것을 하고 있습니다. 드론을 이용해서 사진을 찍고, 드론을 이용해서 영상을 찍고, 사진을 찍은 것을 가지고 영상을 짜서 편집(편집에 관한 특허를 출원중에 있습니다.)해서 책으로 만들어서 판매하는 사업도 하고 있고, 드론을 활용해서 사진촬영을 하는 업체를 발굴해서 홍보를 해서 홍보비를 받는 사업을 하고 있습니다."드론에 관한 특허는 드론이 활용할 수 있는 분야가 대단히 넓다는 것을 보여준다""드론과 관련된 특허는 드론 관련분야의 활용도가 높다는 것을 보여주는 것이다."<|endoftext|>

## 4.3. 사전학습

특허 정보들을 인지했지만, 튜닝 전이라 문장 생성이 어색하다. 일반 질문을 했을 때의 답변에서도 특허와 관련된 내용을 보인다.

---

#### 4.3.1. 일반 질문

##### 1) 건강하게 사는 법

Q: 건강하게 살 수 있는 법 세 가지는?

A: 1.육고기류를 먹지 않고 식물성 단백질을 섭취하는 것.2.육고기류등을 먹지 않는 것.3.고기류를 먹는 것.=====ㄱㄴ 앞의 3가지는 건강을 위해서 필요하지 않는 것이며, 세가지 경우는 건강에 도움이 되는 것이다.1.육고기를 먹지 않는 경우는 먹는 빈도수가 많으면 성인병이 많이생긴다.2.육류를 먹지않으면 성인병인 암,당뇨,고혈압의 원인이 된다.3.육류를먹지않으면 다이어트가 된다.ㄱㄴ"이다.발명(고안)의 국문명칭은 '비만 또는 대사성 질환의 예방, 개선 및 치료용 조성물'이다.발 명(고안)자】, '【성명】', '정태섭', '【특허고객번호】', '4-1995-129099-4'이다.【출원인】', '○【성명】', '윤명섭', '○'이다.

##### 2) 낙타

Q: 낙타가 뭐야

A: 낙타의등분된한쪽옆에는 소를잡아 끌고갈수있는 먹이통이있다. 낙타는먹이를 먹고싶어도 못먹고있는 실정이다. 이런 문제점을 해결하기위하여 1.본발명의 낙타를소에 묶어주는 낙타줄,2.낙타의 목에 줄을 묶어주는 목줄, 3.이줄을 감고 당기는 당김줄, 4.낙타 등에 탈수있게 하는 쇠고리, 5.소에게 먹이를 공급할수있는 구멍,6.소에게 먹이통을 줄수있는 구멍이있는 쇠고리, 7.쇠고리가 빠지지않게 하는쇠고리, 8.쇠고리를 연결할수있는 쇠고리로 구성한것이다."이다.발명(고안)의 국문명칭은 '친환경 태양광 발전구조물'이다.발 명(고안)자의 영문명칭은 'Eco-Friendly Solar Energy Generation Arm', 'Within'이다. 발명(고안자)의 요약내용

#### 4.3.2. 특허 질문

##### 1) 노즐 특허 추천

Q: 노즐에 관한 특허 추천

A: 위의 내용이 있습니다. : 발명자는 이 발명을 하기 위하여 여러사람을 연구하고 노력하였다. 먼저 발명한 사람은 발명의 신기술을 개발하여 발표한 사람이다. 본 발명은 발명의 신 기술을 개발하여 특허출원하였다. 이 발명은 새로운 제품개발이 아니라



---

기존의 제품의 결함을 보완하여 새롭게 개발한 기술이다. 기존의 발명은 기존의 발명은 모두 같은 제품의 명칭과 기술이지만 발명자가 다르기에 특허명칭의 명칭 또한 다르게 되어있으며, 발명의 명칭 또한 다르게 되어 있다. 본 발명의 명칭은 "누설방지기능이 강화된 노즐"이다.[색인어]1. 발명의 명칭누설방지기능을 강화한 노즐.2. 발명의 명칭을 자세히 살펴보면 기존의 발명은 여러사람이 개발을 하였지만 본 발명은 혼자서 연구 개발하여 특허를 출원한 기술이다."이다. 발명(고안)의 국문명칭은 '습도 조절 장치'이다. 발명(고안)자의 성명은 'Humidity control apparatus'

## 2) 자동차 특허

i. Q: 자동차에 관한 특허를 내려고 하는데, 이미 뭐가 있어?

A: 1.본 발명은 자동차용 엔진마운트에 관한 것으로서, 종래에는 엔진마운트의 부품들을 하나씩 교체해야하는 번거로움이 있었고, 또한 작업현장에서 진동에너지를 흡수하는 부품의 구조가 복잡하여 작업성이 저하되는 문제점이 있었다.따라서 본 발명은 엔진마운트의 구조를 개선하여 진동에너지를 흡수하고 동시에 부품들의 내구성을 향상시키고, 또한 운전자가 원하는 부위에 마운트를 고정시킬 수 있도록 하는 자동차용 엔진마운팅 구조에 관한 것으로서 종래에는 여러 개의 마운트를 엔진 마운트에 별도의 브라켓을 사용하여 고정시키는 (후략)

## 4.4. 사전학습 + 파인튜닝

일반 질문에 대하여 비교적 매끄럽게 답변이 가능하다. 특허 질문에서도 전문적인 내용으로 답변이 가능하다.

### 4.4.1. 일반 질문

#### 1) 낙타

Q: 낙타가 뭐야?

A: 낙타는 사막이나 사막 부근에서 사는 동물이고 사막지대의 사막 기후를 이용하기 위하여 낙타와 같이 사막으로 이동할 때에는 뜨거운 사막의 열기를 이용하고 낙타는 뜨거운 사막의 추위를 이용하여 시원한 장소를 찾는다. 사막에서는 태양열의 이용이 어려워,

---

뜨거운 불을 이용해서 바람을 이용하는 것이다. 우리가 겨울에 눈을 쓸 때에 따뜻하게 쓸수 있는 것도 뜨거운 열을 이용하여 눈을 쓸고 나서 따뜻한 열기를 이용해서 눈을 녹인다. 뜨거운물은 몸안의 열을 빼앗아 가고 차가운 물은 몸속의 열을 뺏아가므로 (후략)

## 2) 건강하게 사는 법

Q: 건강하게 사는 법 세 가지는?

A: :1.첫째, 건강하게 사는 법 세 가지를 꼽자면, 첫 번째는 걷기운동을 하며, 둘째는 건강 식품을 먹으며, 셋째는 산소를 마심으로써 몸 건강을 유지한다. 걷기운동은 남녀노소 모두가 일상생활에서 손쉽게 할 수 있어, 누구나 건강관리에 큰 도움을 줄 수 있다.또한, 두 번째는 산소를 마시는 것으로, 산소를 마시는 행위는 자연과 인간의 조화로운삶을 추구하는 것이다. 즉, 자연과 인간은 함께 살고 싶어하지만, 함께 살 수 없는 서로다른 존재이다. 그래서 인간은 자연의 일부인 산소를 마시기 위하여, 인간의 의지에 의한 호흡과정을 거치면서 자연과 인간은 인간의 의지와는 상관없이 인간과 자연이 함께 공존할 수 없는 관계가 되었다. (후략)

### 4.4.2. 특허 질문

#### 1) 노즐 특허

Q: 노즐에 관한 특허

A: 특허정보넷의 특허공보를 참고하시면 노즐 관련 특허를 찾을 수 있습니다. 특허공보에서는 각 분야별 특허를 제시하고 있으므로 해당 분야의 특허를 검색하여 찾아보시면 도움이 될 것입니다. 특허공보에는 이 분야의 특허도 포함되어 있습니다. 더 자세한 내용은 한국기술산업(주)의 특허자료로 확인하시면 좋습니다. 더 자세하고 도움이 필요하신 경우, 해당 분야의 특허정보를 검색해보시기 바랍니다. (전화번호: 02-202-7731) 감사합니다.

추가로 궁금한 사항이 있으시면 경기지방중소벤처기업청 비즈니스지원단에 문의하시면 됩니다. (전화번호 : 02-202 - 7731) 감사합니다."이다. 추가로 문의하실 사항이 있으시다면 한국기업정보(주)로 문의하시면 도움을 받을 수 있습니다."라고 하였습니다. 감사한 내용이 있을 것으로 예상되며, 추가로 도움이 필요한 경우 더 문의할 수 있다는 말을 포함하였습니다. 추가적인 도움이 필요한 내용을 문의할 수 있는 곳으로 한국

---

## 2) 자동차 특허

Q: 내가 자동차에 관한 특허를 만들었는데, 혹시 어떤 특허들이 있어?

A: 1. 제1 특허: 자동차 관련 산업에 대한 기술적 내용을 기술합니다. 이 특허는 자동차의 작동 방법과 관련된 기술을 포함합니다.

2. 제2 특허: 자동차 관련 기술이 산업에 적용된 예를 기술합니다. 이는 자동차의 다양한 형태와 관련된 기술을 말합니다.

3. 제3 특허: 자동차의 산업적 활용에 대한 기술을 포함합니다. 이는 자동차의 작동 방식, 내부 공간 활용 등과 관련된 산업적 이용을 말합니다. 이러한 기술들은 자동차의 품질 향상에 도움을 주고, 경제 활성화에 기여할 수 있습니다.

(중략)

이러한 특허들은 자동차의 산업적 이용에 대한 기술입니다. 이와 같은 기술들은 자동차 산업의 경제와 산업적 이용의 촉진에 기여할 수가 있습니다. 이러한 기술은 자동차의 품질과 환경에 대한 개선을 가져올 수 있으며, 경제 활성화와 환경 문제 개선에 기여할 수 있습니다.

## 2) 드론 특허

Q: 드론에 관한 특허에는 뭐가 있어?

A: 기존 드론에 관한 특허는 드론의 모터의 회전 수를 조절하거나, 또는 모터의 회전수를 조절하지 않고 드론의 추락에 대한 위험을 방지하는 안전성이 있는 반면, 드론의 추락에 의한 안전사고의 위험성이 있었다. 본 발명은 기존 드론의 안전사고와 대비하여 비행시간을 증가시키고, 비행의 안정성을 높인 비행을 위한 것이다. 기존 드론에 비행안전을 위한 안전장치를 부착하고, 기존 드론보다 더 큰 모터를 장착하여 비행의 안정성, 속도 및 비행거리를 증가시킬 수 있고, 추락시 더 큰 회전 수를 만들어 추락하는 드론을 보호하게 된다."이다. 발명(고안)의 국문명칭은 '드론의 안전착륙 방법'이다. 발명(고안)자의 국문명칭은 '드론의 안전 착륙을 위한 무선 충전 시스템 및 그 동작 방법'이다."이다. 발명(고안)자의 국문명칭은 지면에 고정된 지상 통제 장치에 대한 안전 거리를 확보하는 방법을 포함하', '는 드론 시스템 및 그 제어 방법'이다. 발명(고안)의 영문명칭은 ' (후략)

#### 4.5. 연구 결과 분석

1. 기존모델: 한글 생성은 잘 하지만 문장 구성이 잘 안 됨.
2. 기존모델 + 파인튜닝: 기본 문답은 가능하지만 특허 질문을 처리하지 못함
3. 사전학습: 특허 내용은 인지했지만 문장 생성이 다소 어색함
4. 사전학습 + 파인튜닝: 일반, 특허 질문 모두 답변 가능함

	일반질문	특허질문	문장구성
기존모델	질문에 답하지 않음	전혀 무관한 답을 보임	문장 구성이 올바르지 않아 하고자 하는 말을 이해하기 힘들
기존모델 + 파인튜닝	질문에 적합한 답변	특허라는 단어를 쓰기는 하지만 특허에 관해 알려주지는 않음	기존모델만 사용했을 때보다는 낫다. 문장 간의 연결이 매끄럽지 않음
사전학습	질문에 적합하지 않은 답변, 특정 도메인(특허)에 관한 내용이 언급됨	특허에 관해 답하는 듯 하지만 다소 두루뭉술하고 특허의 이름 등 명확한 정보를 주지 않음	문장 간의 연결이 매끄럽지 않음
사전학습 + 파인튜닝	질문에 대해 비교적 적합한 답변을 보임	정제된 답변을 보여줌, 특허에 관해 언급함	비교적 매끄럽고 이해하기에 간결한 문장을 구성함

### 5. 결론 및 향후 연구 방향

#### 5.1. 결론

사전학습 과정에서 ai모델이 대량의 텍스트 데이터를 학습하여 일반적인 언어 패턴과 지

---

식을 습득하였고 파인튜닝 과정에서는 특정 도메인에 적합한 고유의 특성과 패턴을 학습하였음을 알 수 있다. 이러한 과정을 통해 최적화된 모델은 매끄러운 문장 생성 능력을 향상시키고, 도메인에 특화된 작업에서의 성능을 개선하였음을 확인할 수 있었다.

## 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 데이터를 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](http://www.kipris.or.kr/khome/main.jsp)

## 2023 년 전기 산학협력 프로젝트 멘토 의견서

### 1. 지도개요

팀 명	응답하라 챗봇		
과 제 명	도메인 지식에 특화된 한국어 대규모 언어모형 기반 챗봇		
협력기관	(주)나라인포테크		
참여학생	이름	전화번호	이메일
	김정민	010-9210-6989	kjmin017@gmail.com
	김채원	010-2798-2765	bbubbune@naver.com
	최상준	010-8523-0874	choi4th4570@gmail.com
참여교수명	권혁철		

### 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.

위 내용을 부산대학교 정보컴퓨터공학부 2023 학년도 전기 산학협력프로젝트 지도내용으로 제출합니다.

멘토링 일시	2023 년 09 월 21 일	시작시간	15:00:	종료시간	17:00
--------	------------------	------	--------	------	-------

소속: (주)나라인포테크

직급: 연구소장

성명: 이승희

(서명)