

KorBert 모델을 활용한 맥주 추천 챗봇 개발



Work Team & Members

3 팀 : 4김1박

팀장 : 김세리 (PL/QA)

팀원 : 김다정(DA), 김예지(DA), 김용관(DA), 박재형(TA/AA)



Work schedule

22.05.30 프로젝트 시작

22.06.03 데이터 구축

22.06.10 모델 학습 및 평가

22.06.17 챗봇 구현

22.06.24 프로젝트 종료



Work Rule

DataSet : 크롤링을 활용하여 데이터 수집

Model : ETRI KorBERT

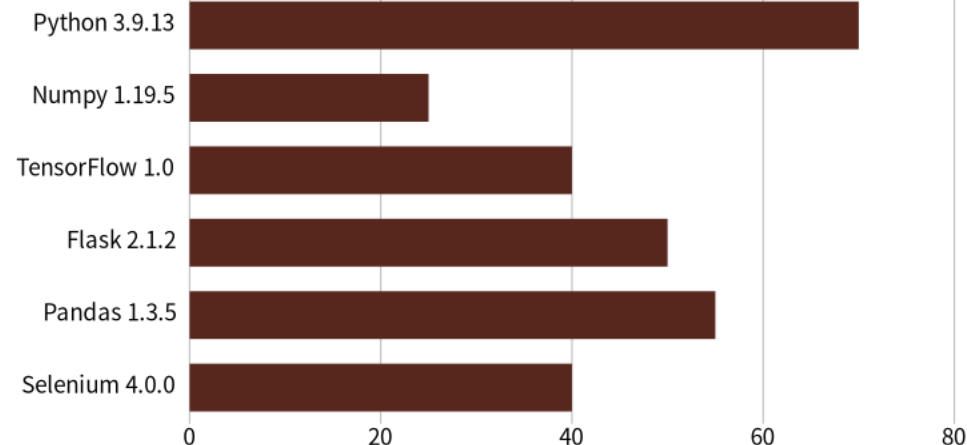


KorBert를 활용한 맥주 추천 챗봇 개발

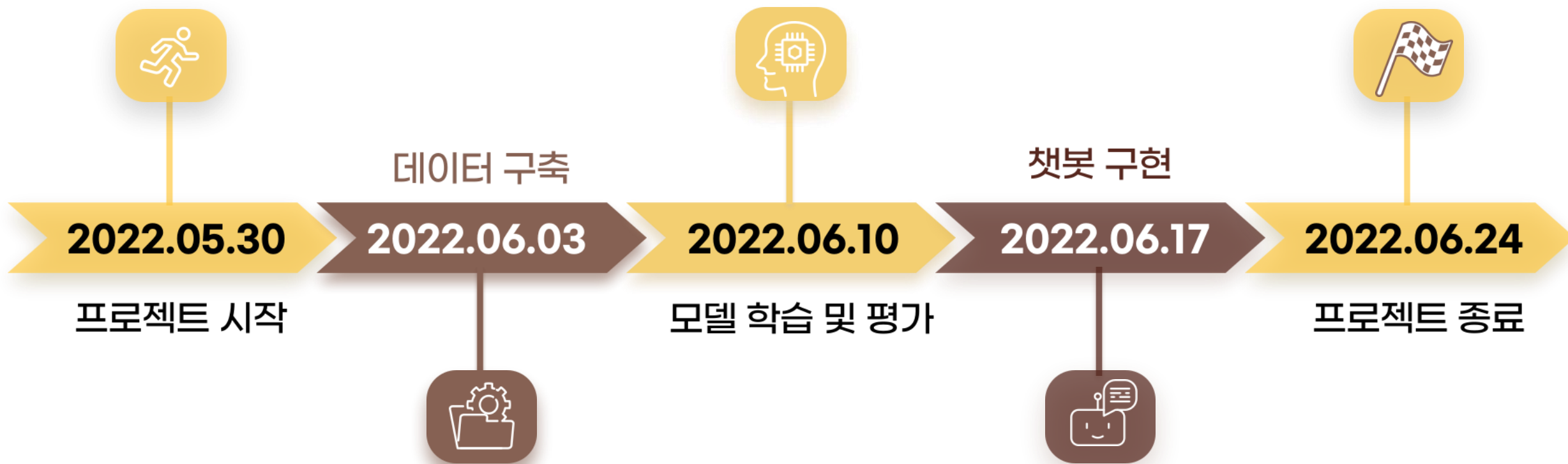
ETRI KorBERT를 활용하여 생활맥주의 맥주 메뉴 데이터를 크롤링하여 메뉴명, 맥주 종류, 맛과 향으로 키워드를 분류하여 사용자의 취향에 따른 맥주를 추천해주는 챗봇 개발



Skills



TIMELINE





4김1박 팀원 소개



PL/QA
김세리

- ▶ 기획안 작성
- ▶ 학습용 데이터 문장 생성
- ▶ 생성된 문장 예외 처리 & 예외 처리 코드 구현
- ▶ 추천 알고리즘 구현 총괄
- ▶ 알고리즘 예외 처리 코드 구현
- ▶ 코드 병합 총괄
- ▶ 산출물 품질 관리
- ▶ PPT 작성



DA
김예지

- ▶ 기획안 작성
- ▶ 학습용 데이터 문장 생성
- ▶ 생성된 문장 예외 처리
- ▶ 크롤링 데이터 전처리
- ▶ PPT 작성



DA
김용관

- ▶ 기획안 작성
- ▶ 텍스트 & 이미지 크롤링
- ▶ 학습용 데이터 문장 생성
- ▶ 생성된 문장 예외 처리
- ▶ 이미지 출력 함수 구현
- ▶ HTML & CSS 파일 수정
- ▶ PPT 작성



DA
김다정


- ▶ 기획안 작성
- ▶ 크롤링 데이터 전처리
- ▶ 학습용 데이터 문장 생성
- ▶ 생성된 문장 예외 처리
- ▶ 추천 알고리즘 기본 틀 구현
- ▶ 알고리즘 예외 처리 코드 구현
- ▶ HTML & CSS 파일 수정
- ▶ PPT 작성



TA/AA
박재형

- ▶ 기획안 작성
- ▶ 이미지 크롤링
- ▶ 학습용 데이터 문장 생성
- ▶ 생성된 문장 예외 처리 & 예외 처리 코드 구현
- ▶ 추천 알고리즘 구현 총괄
- ▶ 알고리즘 예외 처리 코드 구현
- ▶ 코드 병합
- ▶ PPT 총괄

INDEX

- 
- 01. 챗봇 추진 배경
 - 02. 챗봇 추진 방안
 - 03. 챗봇 구현
 - 04. 기대 효과
 - 05. 소감 및 참고문헌

01. 챗봇 추진 배경





1.1 기획 배경 1

1. 챗봇 추진 배경

- ▶ 최근 몇 년간 수제 맥주에 대한 관심이 높아져서 시장 규모가 급성장했음

[그림4-1] 국내 수제 맥주 시장규모 및 제조업체 현황
<수제 맥주 시장규모> (단위 : 억 원) <수제 맥주 제조업체 현황> (단위 : 개)



* 2021년은 상반기 기준 현황
출처 : 한국수제맥주협회

2일 한국수제맥주협회에 따르면 지난해 국내 맥주시장에서 국산 수제맥주 규모는 1천180억원으로 나타났다.

2017년 433억원과 비교하면 3년 만에 2.7배 성장한 것이다. 수제맥주협회는 2023년에는 시장 규모가 3천700억원까지 성장할 것으로 추정했다.

한국조세재정연구원의 발표에 따르면 국산 수제맥주 시장 규모는 혼술, 홈술 등 다양한 맛을 추구하는 소비자 증가로 2020년 1천180억 원으로 3년 전인 2017년보다 173%나 급성장했다. 2023년에는 3천700억 원 규모로 성장할 전망이다.

- ▶ 2020년 기준, 국내 맥주 시장 규모 5조 중 수제 맥주 시장 규모는 1,200억으로 비교적 작은 규모지만 5년 사이에 약 5배 가량 증가함
- ▶ 제조업체 또한 매년 증가하고 있는 추세



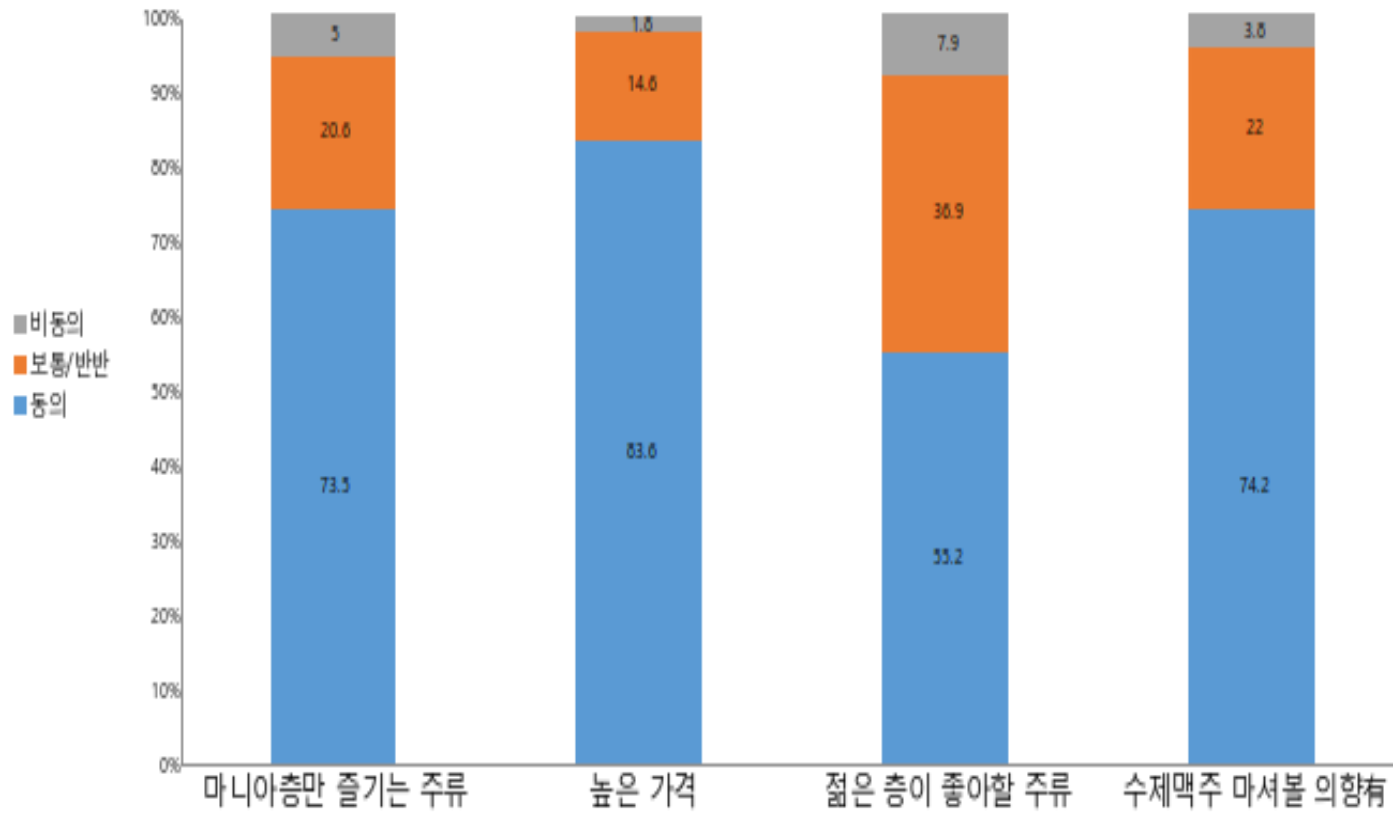


1.1 기획 배경 2

1. 챗봇 추진 배경

- ▶ 그렇지만 아직 수제 맥주에 대한 정보가 널리 알려지지 않아 수제 맥주를 접해본 사람은 많지 않음

수제 맥주 관련 전반적 인식 평가



- ▶ 수제맥주를 경험한 비율은 약 전체 중 18%
- ▶ 수제 맥주에 대한 경험은 부족하지만 경험했을 때 긍정적인 평가
- ▶ 수제 맥주에 대한 정보와 판매처가 부족하고 활성화 되어있지 않음
- ▶ 소비자들이 경험했을 때 긍정적인 평가가 높고 향후 경험에 대한 긍정적 인식과 의향이 있는 반면 정보와 판매처가 따라가지 못하는 상황



1.2 추진 목표

1. 챗봇 추진 배경

- ▶ '생활맥주'의 맥주 추천 챗봇을 개발하여 소비자의 접근성을 높여 수제 맥주의 대중화에 기여하고자 함



- ▶ 생활맥주는 전국 각지의 양조장과 협업을 통해 차별화된 수제맥주를 선보이는 수제맥주 플랫폼
- ▶ 유통채널이 없는 지역 양조장의 수제맥주를 유통해 발전시킴
- ▶ 현재 '생활맥주'는 국내 수제맥주 대표 브랜드로써, 수제맥주를 경험하기 가장 좋은 브랜드라고 판단
- ▶ 수제맥주를 처음 즐기는 소비자들에게 소비자 입맛에 맞는 맥주를 추천하는 챗봇을 제공함으로써 수제맥주에 대한 긍정적 인식과 평가 기대

사용자가 찾는 정보를 바탕으로 수제 맥주를 추천해 주는 챗봇 서비스를 기획

02. 챗봇 추진 방안





2.1 개발 환경 선정

2. 챗봇 추진 방안

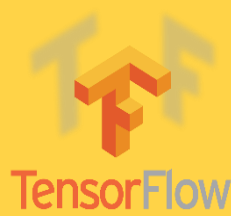
개발 환경



웹 디자인



알고리즘



BeautifulSoup

ETRI



2.2 훈련 데이터셋 수집 및 전처리

2. 챗봇 추진 방안

- ▶ 크롤링을 통한 데이터 수집, 문장들 생성 후 문장 확대 및 예외처리를 통해 훈련 데이터셋의 정확도를 높임



데이터 수집



문장들 생성



예외처리

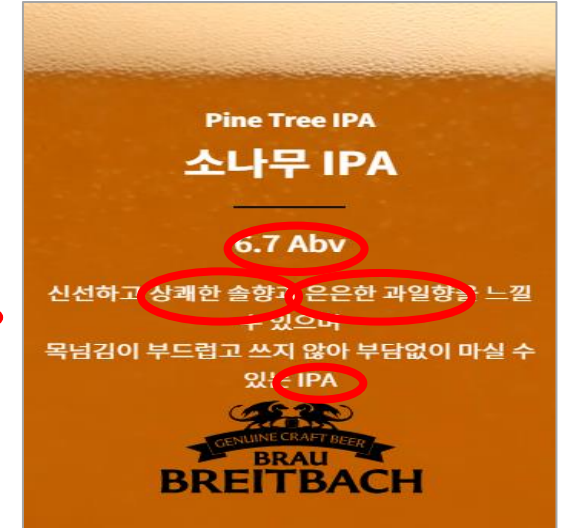


2.2 훈련 데이터셋 수집 및 전처리

2. 챗봇 추진 방안

- ▶ 크롤링으로 데이터 수집 후, 이를 바탕으로 “종류, 도수, 맛, 향” 이라는 4가지의 맥주 정보를 슬롯으로 선정

- ▶ 생활맥주에서 판매중인 수제 맥주에 대한 이미지와 텍스트 크롤링
- ▶ 수집한 맥주의 정보에는 ‘종류’, ‘도수’, ‘향’, ‘맛’이 공통적으로 들어가 있음
- ▶ 공통된 정보를 슬롯으로 설정
- ▶ 정보를 바탕으로 데이터 전처리



eng_name	kor_name	abv	types	flavor	taste	info
American	태평양 IPA	7.1	IPA	상큼한 홉	쓴, 씹쓸한	미국 서부 정통 IPA로 홉의 화사하고 달콤 상큼한 향미와
National t	국보 IPA	6.9	IPA	꽃, 상큼한	안단, 안단	국산곡물(김해쌀, 군산맥아, 진주얇은뱅이밀)과 국산홉을
Pine Tree	소나무 IPA	6.7	IPA	상큼한, 과일	안단, 안단	신선하고 상쾌한 솔향과 은은한 과일향을 느낄 수 있으며
Neuer	노이어	6.5	에일	상큼한, 과일	안단, 안단	파인애플, 망고, 자몽을 넣은 듯한 시트러스한 아로마와
American	대청봉 IPA	6.4	IPA	상큼한, 과일	안단, 안단	부드러운 카라멜이 연상되는 달콤, 고소한 맛과 상큼한 과
American	겉작 IPA	6.3	IPA	꽃, 홉	쓴, 씹쓸한	목직한 홉의 씹새롭함과 꽃향기가 섬세하게 피어오르는
American	놀라운 IPA	6.2	IPA	과일	쓴, 씹쓸한	미국 서부식 IPA로, 오렌지, 자몽 류의 풍부한 과일 캐릭터
Red IPA	고릴라 RE	6.2	IPA	상큼한, 홉	단, 달달한	영국산 최고급 레드 몰트의 달콤한 풍미와 국내산 생홉의
Baristout	바리스타우	6	흑맥주	커피, 스모크	쓴, 씹쓸한	4종의 원두를 블렌딩해 넣어 바리스타가 갓 내린 진한 커피
Sour Ale	셀레임 플	6	에일	상큼한, 과일	신, 상큼한	젖산 발효의 새콤한 신맛과 상큼한 꿀 껌질향의 홉이 조화
American	강남 페일	5.9	에일	과일, 홉	단, 달달한	폭발하는 열대 과실향의 홉 풍미와 달달한 캐러멜 몰트가
Saison	불란서 로	5.9	에일	상큼한	신, 상큼한	프레쉬한 산미와 드라이한 피니쉬로 갈증해소에 적합한 프
American	런던 IPA	5.8	IPA	홉	단, 달달한	감미롭게 채워지는 달콤한 몰트 풍미와 화사한 홉의 향미,
American	일상탈출 I	5.8	IPA	상큼한, 과일	안단, 안단	부드럽게 채워지는 영국 효모 특유의 풍미와 열대 과일향
Sopping II	홉백 IPA	5.7	IPA	상큼한, 과일	쓴, 씹쓸한	화사한 자몽, 오렌지향과 더불어 알싸한 향신료 향을 통해
Lager	페스트 비	5.5	라거	꽃, 홉	단, 달달한	구운 빵의 고소함과 달달한 꿀 향의 조화! 독일 정통식 최
Camellia	동백밀맥주	5.3	바이젠	상큼한, 과일	단, 달달한	동백의 도시 통영에서 온 밀맥주로 부드러운 목넘김, 상쾌



2.2 훈련 데이터셋 수집 및 전처리

2. 챗봇 추진 방안

- ▶ 크롤링으로 데이터 수집 후, 이를 바탕으로 “종류, 도수, 맛, 향” 이라는 4가지의 맥주 정보를 슬롯으로 선정

- ▶ 모델이 학습할 문장 틀을 제작
- ▶ 문장에 슬롯이 포함되어 있지 않은 문장부터 모든 슬롯을 포함한 문장 틀 생성
- ▶ 틀을 바탕으로 문장에 토큰값을 바꿔주면서 약 10,000 개의 문장을 확대 생성

```
# 슬롯 : 종류, 도수, 맛, 향
beer_types = ['/types;메일/', '/types;IPA/', '/types;라거/', '/types;바이젠/', '/types;흑맥주/']

beer_abv = ['/abv;3도/', '/abv;4도/', '/abv;5도/', '/abv;6도/', '/abv;7도/', '/abv;8도/',
            '/abv;3도이상/', '/abv;4도이상/', '/abv;5도이상/', '/abv;6도이상/', '/abv;7도이상/',
            '/abv;3도 이하/', '/abv;4도 이하/', '/abv;5도 이하/', '/abv;6도 이하/', '/abv;7도 이하/',
            '/abv;4도이하/', '/abv;5도이하/', '/abv;6도이하/', '/abv;7도이하/', '/abv;8도이하/',
            '/abv;4도 이하/', '/abv;5도 이하/', '/abv;6도 이하/', '/abv;7도 이하/', '/abv;8도 이하/']

beer_flavor = ['/flavor;과일/', '/flavor;홉/', '/flavor;꽃/', '/flavor;상큼한/', '/flavor;커피/', '/flavor;스모키한/']

beer_taste = ['/taste;단/', '/taste;달달한/', '/taste;달콤한/', '/taste;안단/', '/taste;안 단/',
              '/taste;달지 않은/', '/taste;달지않은/', '/taste;쓴/', '/taste;쓴쓸한/',
              '/taste;짭짭한/', '/taste;달콤쓴쓸한/', '/taste;안쓴/', '/taste;안 쓴/', '/taste;쓰지 않은/',
              '/taste;신/', '/taste;상큼한/', '/taste;새콤달콤한/', '/taste;시지 않은/', '/taste;시지않은/',
              '/taste;쓰지않은/', '/taste;안신/', '/taste;안 신/', '/taste;과일/', '/taste;고소한/', '/taste;구수한/']
```

```
# 종류 도수 향 맛
{1} {2} {3} {4} 걸로
종류 {1} 도수 {2} 향 {3}향 맛 {4}맛
{1}에 {2} {3}향나는 {4}맛 맥주
{1}중에 {2}이고 향은 {3}향인 {4} 맥주 추천
종류는 {1}인데 {2}에 {3}냄새가 나는 {4}맛 맥주 찾아줘
{1}에 {2}이고 {3}냄새인 {4} 거 있어?

# 종류 도수 맛 향
{1} {2} {4} {3}
{1} {2}고 {4}맛 {3}향인 거 추천해줘
{1} {2}이고 {4}맛 {3}향인 맥주 찾아줘
{1}중에 {2} {4}맛이 나면서 {3}냄새가 나는 걸로
```




2.2 훈련 데이터셋 수집 및 전처리

2. 챗봇 추진 방안

▶ 다양한 문장들을 제작 후 틀을 바탕으로 약 10,000개의 문장을 확대 생성한 예시

```
# 종류 도수 향 맛
{1} {2} {3} {4} 걸로
종류 {1} 도수 {2} 향 {3}향 맛 {4}맛
{1}에 {2} {3}향나는 {4}맛 맥주
{1}중에 {2}이고 향은 {3}향인 {4} 맥주 추천
종류는 {1}인데 {2}에 {3}냄새가 나는 {4}맛 맥주 찾아줘
{1}에 {2}이고 {3}냄새인 {4} 거 있어?
```

```
# 종류 도수 맛 향
{1} {2} {4} {3}
{1} {2}고 {4}맛 {3}향인 거 추천해줘
{1} {2}이고 {4}맛 {3}향인 맥주 찾아줘
{1}중에 {2} {4}맛이 나면서 {3}냄새가 나는 걸로
```

```
# 종류 향 도수 맛
{1} {3}거 {2} {4} 맥주
{1}이고 {3}향 {2} {4}맛
{1} 종류에 {3}냄새 {2} {4} 맥주로
{1}중에서 {3}향 {2} {4} 게 있을까?
{1} {3}향이 나고 {2}에 {4}맛인 맥주 추천해주세요
```

문장 틀 생성

문장들
확대

```
/abv;4도/이고 /types;IPA/에 /taste;쓰지 않은/맛 /flavor;과일/냄새인 걸로 찾아줘
/abv;7도/ 맥주 그리고 /flavor;과일/향이 있으면 좋겠어
/abv;8도이하/ /types;라거/ /flavor;과일/ /taste;달지않은/ 거
/flavor;꽃/향 나고 /abv;6도 이하/ 맥주
/taste;썩썩한/ /types;흑맥주/
종류는 /types;흑맥주/이고 /abv;7도 이상/ 도수의 /taste;새콤달콤한/ 맛 맥주
나는 /abv;3도 이상/ 맥주가 좋아
/abv;7도 이하/인 /types;라거/ 중에 /flavor;홉/향이 나는 맥주 추천해 주세요.
/types;에일/ 중에 /abv;3도이상/인 맥주 추천해주세요.
/taste;안단/ 건데 /flavor;상큼한/향도 나는 /types;라거/ 맥주
/taste;안 신/ /types;에일/ /abv;4도 이하/ /flavor;꽃/
/abv;3도/ /types;IPA/ /taste;안신/ /flavor;과일/맥주
/abv;4도/ /types;바이젠/ 좋아해
/flavor;꽃/ /abv;5도/ /types;에일/ /taste;새콤달콤한/ 맛로
/abv;3도 이상/ 맥주 중에 /taste;쓰지않은/맛이 나는 맥주 추천해주세요.
/abv;7도이상/에 /taste;안 쓴/맛 나는 거 추천해줘
```

문장 틀 확대



2.2 훈련 데이터셋 수집 및 전처리

2. 챗봇 추진 방안

▶ 부자연스럽게 생성된 문장과 단어 예외처리

- ▶ 슬롯에 맞게 문장을 만들었지만 슬롯에 들어간 토큰값이 품사, 받침, 띄어쓰기 등의 문제로 부자연스러운 문장들이 생성
- ▶ 문장을 자연스럽게 만들기 위해 정규표현식으로 예외처리
- ▶ 부자연스러운 문장을 처리함으로써 더 나은 모델 학습을 할 수 있도록 함
- ▶ 정보를 바탕으로 데이터 전처리

```
#if (ord(rnd_types[-2])-ord('가')) % 28 == 4 or 8: # 4면 받침이 나, 8이면 ㄴ
if 44032 <= ord(rnd_types[-2]) <= 55203 and (ord(rnd_types[-2]) - ord("가")) % 28 > 0:
    line = line.replace(rnd_types+'면', rnd_types+'미면')
    line = line.replace(rnd_types+'가', rnd_types+'미')
    line = line.replace(rnd_types+'였으면', rnd_types+'미였으면')

if (ord(rnd_flavor[-2])-ord('가')) % 28 != 4 :
    for reg in ['거로', '거', '컬로', '게', '컨데', '으로', '것', '거로', '거', '컬로', '게', '컨데', '으로', '것']:
        line = line.replace(rnd_flavor+reg, rnd_flavor+'항 나는'+reg)

for regex in ['으로', '으로']:
    line = line.replace(rnd_flavor+regex, rnd_flavor+'항으로')

if (ord(rnd_flavor[-2])-ord('가')) % 28 == 4:
    line = line.replace(rnd_flavor+'로', rnd_flavor+'으로')

# 정규 표현식
line = re.sub((r'/types;흑맥주/\s*맥주\s*'), '/types;흑맥주/', line)
line = re.sub((r'/types;바이젠/\s*로'), '/types;바이젠/으로', line)

line = re.sub((r'도\s*이상/로'), '도 이상/으로', line)

line = re.sub((r'/taste;과일/\s*(거|컬로|게|컨데|으로|것)'), '/taste;과일/맛 나는 거', line)

slots.append(line)
```



2.2 훈련 데이터셋 수집 및 전처리

2. 챗봇 추진 방안

▶ 부자연스럽게 생성된 문장과 단어 예외처리 예시

/taste;신/맛이랑 /flavor;상큼한/향이 나는 /types;바이젠/로 추천해줘
/abv;8도/이면서 /taste;고소한/ /types;흑맥주/ 맥주로 추천해줘.
냄새는 /flavor;과일/향 거 맛은 /taste;단/거 종류는 /types;에일/ 도수는 /abv;6도 이상/로
/types;바이젠/였으면 좋겠고 /taste;달콤한/ 거
/types;라거/에 냄새는 /flavor;커피/향 /taste;안쓰/맛이고 도수는 /abv;5도 /으로!
/flavor;꽃/으로

/taste;신/맛이랑 /flavor;상큼한/향이 나는 /types;바이젠/으로 추천해줘
/abv;8도/이면서 /taste;고소한/ /types;흑맥주/ 로 추천해줘.
냄새는 /flavor;과일/향 나는거 맛은 /taste;단/거 종류는 /types;에일/ 도수는 /abv;6도 이상/으로
/types;바이젠/이었으면 좋겠고 /taste;달콤한/ 거
/types;라거/에 냄새는 /flavor;커피/향 /taste;안쓰/맛이고 도수는 /abv;5도 이상/으로!
/flavor;꽃/향으로




2.3 학습 모델 탐구 및 선정

2. 챗봇 추진 방안

▶ 챗봇 제작에 필요한 자연어 처리에 적합한 모델을 찾기 위해 논문을 탐색

- ▶ 우리가 만들고자 하는 건 슬롯 채우기(Slot Filling)를 이용한 목적지향 챗봇(Task-Oriented Dialogue System)
- ▶ 맥주 종류의 증감에 따른 슬롯과 시스템 업데이트가 잦을 수 있으므로, 확장성(Scalability)이 떨어지면 슬롯이 업데이트 될 때마다 새로 학습시켜야 하는 문제 발생
- ▶ 기존에 학습된 모델을 활용해서 새로운 데이터 없이도 처리할 수 있는 모델이 필요

- 
- ▶ BERT의 언어처리 능력을 활용하면 한 번도 본 적 없는 데이터도 처리할 수 있는 Zero-shot Learning이 가능
 - ▶ 슬롯이나 서비스가 업데이트 되더라도 새로 학습시킬 필요가 없음, 즉 BERT는 확장성이 높아 슬롯 채기 목적지향 챗봇 개발에 적합



2.3 학습 모델 탐구 및 선정 : BERT

2. 챗봇 추진 방안

- ▶ 게다가 BERT는 다양한 NLP Task에서 좋은 성능을 보였으며, 일부는 사람을 뛰어넘을 만큼 훌륭한 성과를 거둔 자연어 처리 모델이므로, 챗봇 개발에 적합하다고 판단하여 모델로 선정함

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Table 1: GLUE Test results, scored by the evaluation server (<https://gluebenchmark.com/leaderboard>). The number below each task denotes the number of training examples. The “Average” column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.⁸ BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.



2.4 BERT의 학습 과정

2. 챗봇 추진 방안

- ▶ 자연어 처리 분야에서 가장 우수한 성능을 보여준 BERT



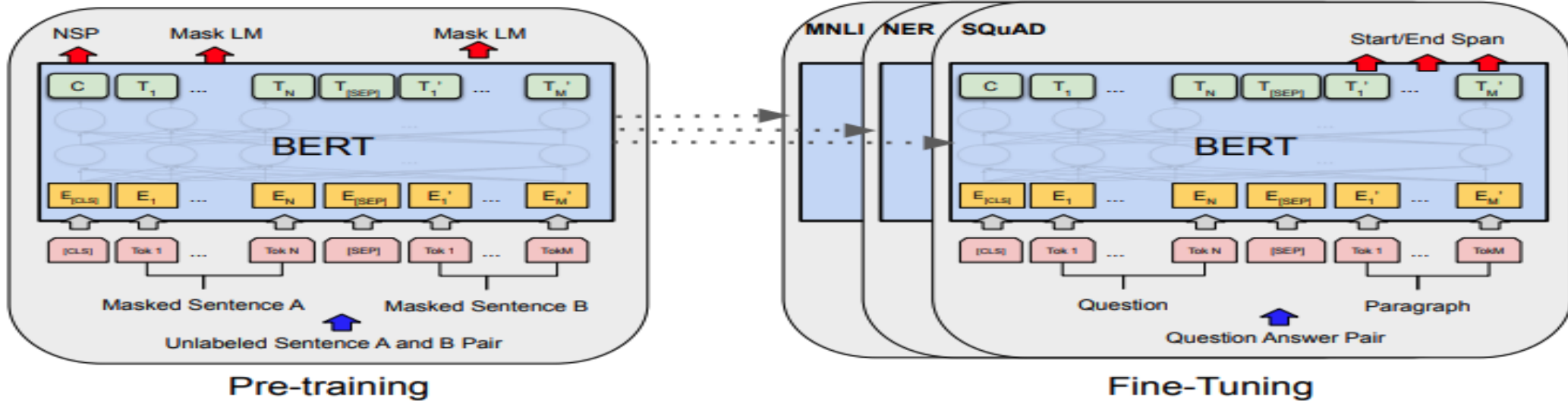
- ▶ 2018년에 구글이 공개한 자연어 처리 딥러닝 모델
- ▶ 트랜스포머의 인코더를 사용하여 사전 훈련된 언어표현 모델
- ▶ 자연어 처리 태스크를 교육 없이 양방향으로 사전학습하는 첫 시스템
- ▶ 기본적으로 대량의 단어 임베딩 등에 대해 사전 학습이 되어 있는 모델을 제공하기 때문에 상대적으로 적은 자원만으로도 좋은 성능을 냄
- ▶ 원하는 문제를 해결하기 위한 fine-tuning만 하면 되기 때문에 훈련 시간이 짧음



2.4 BERT의 학습 과정

2. 챗봇 추진 방안

- ▶ BERT의 자연어 처리는 Pre-training과 Fine-Tuning의 2단계로 진행됨



- ▶ BERT 모델이 대용량 코퍼스를 기반으로 언어 전반에 대해서 이해하는 과정
- ▶ 대용량 코퍼스를 스스로 라벨링하는 self-supervised 학습

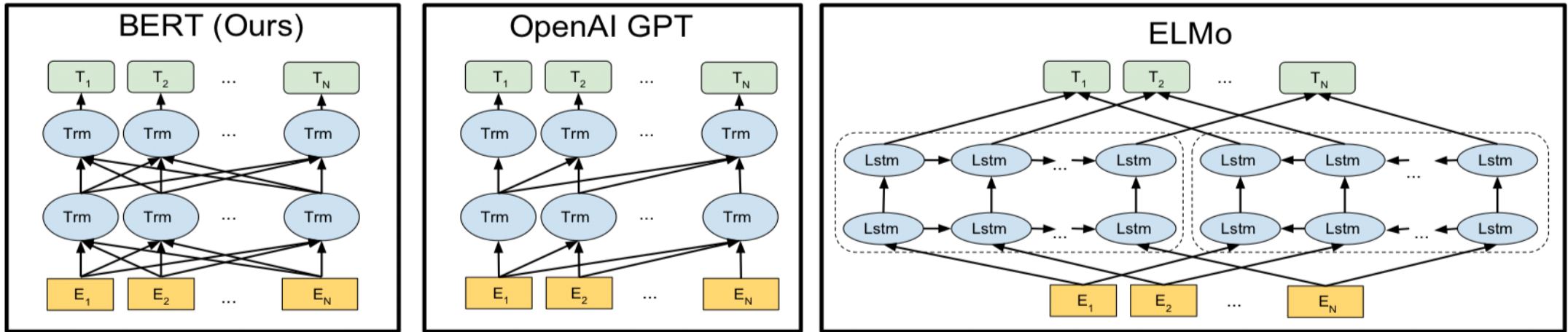
- ▶ 사전학습된 BERT 모델을 기반으로 수행하고자 하는 작업에 알맞게 추가적인 학습을 하는 것
- ▶ 각 작업에 알맞게 준비된 별개의 라벨링된 데이터 必



2.4 BERT의 학습 과정

2. 챗봇 추진 방안

- ▶ BERT의 Pre-Training 방법론은 MLM과 Next Sentence Prediction



- ▶ 방향성이 없거나, 양방향적이지만 그 깊이가 얇은 기존 모델의 단점 극복
- ▶ Masked Language Model(MLM)
- ▶ Next Sentence Prediction



2.4 BERT의 학습 과정

2. 챗봇 추진 방안

- ▶ BERT의 Pre-Training은 토큰으로 문장을 분리하고, 세 종류의 임베딩을 통합

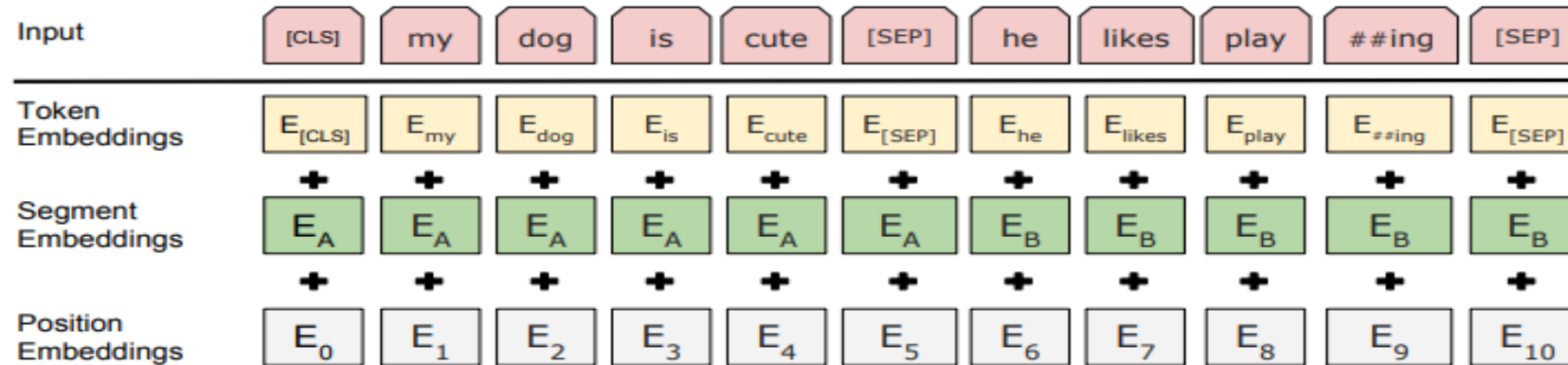


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

- ▶ 토큰 ([SEP], [SEP])을 통해 기존의 문장들을 분리
- ▶ Token Embedding : 실질적인 입력이 되는 단어 임베딩
- ▶ Position Embedding : 위치 정보를 학습하기 위한 임베딩
- ▶ Segment Embedding : 서로 다른 문장을 구분하기 위한 임베딩



2.5 BERT 모델 선정 : KorBERT

2. 챗봇 추진 방안

▶ BERT 모델 중 교착어인 한국어에 최적화된 언어 모델인 ETRI의 KorBERT 사용



Electronics and Telecommunications
Research Institute

KorBERT

▶ 버트에 비해 5가지 성능 평가 항목에서 뛰어남 :

문장 유사도 추론, 주제 분류, 기계독해 등 영역에서 정확도 성능이 평균 4.5% 우수
특히 정답 단락을 순위화 하는 '단락순위화' 평가에서는 7.4% 높았음

▶ 교착어인 한국어 특성을 반영 :

버트는 단어가 모두 떨어져 있는 고립어 기반,
코버트는 어근에 접사가 붙는 교착어 기반.
ETRI는 고립어에 맞게 의미 최소 단위인 형태소까지 고려하는 모델을 구현하여, 정확도를 높임

▶ 학습 데이터 증가 :

버트가 사용한 데이터는 40여만건 위키백과 문서.
ETRI는 여기에 총 23기가바이트(GB)에 달하는
신문기사·백과사전 정보를 더해 45억개 형태소를 학습시킴



2.5 BERT 모델 선정 : KorBERT

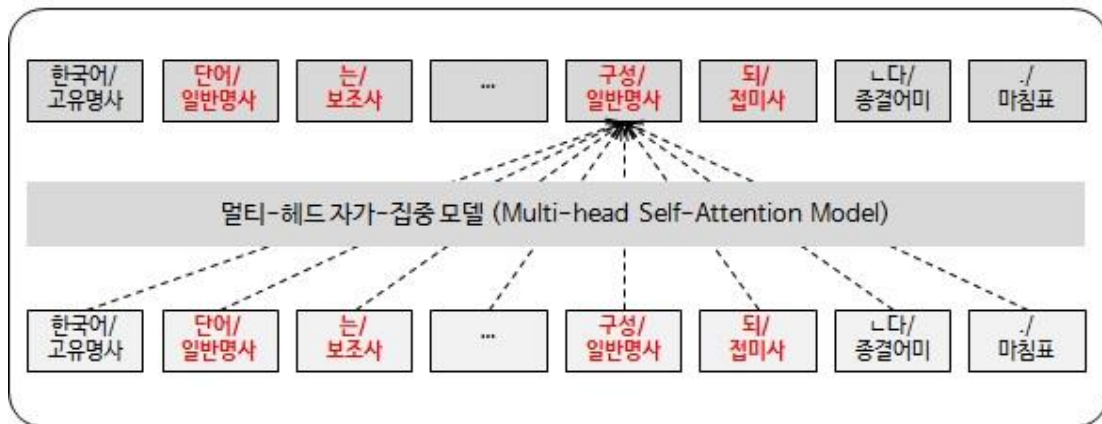
2. 챗봇 추진 방안

▶ 모델에 대한 장단점 또는 모델 테스트 결과

한국어분석 태스크 (의미역인식 성능: 85.77%) 기계독해 태스크 (KorQuAD dev셋: 94.18%) 문장/문서분류 태스크 (단락재순위: 73.73%)

언어모델 단위 별 문맥 반영 워드 벡터

한국어/고유명사 단어/일반명사 는/보조사 ... 구성/일반명사 되/접미사 는다/종결어미 마침표



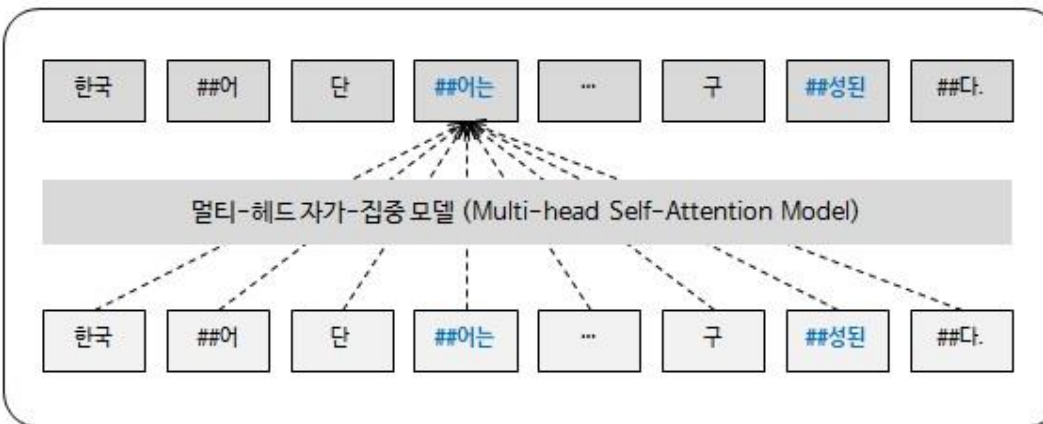
한국어/고유명사 단어/일반명사 는/보조사 ... 구성/일반명사 되/접미사 는다/종결어미 마침표

ETRI 형태소 기반 언어모델 분석 단위

한국어분석 태스크 (의미역인식 성능: 81.85%) 기계독해 태스크 (KorQuAD dev셋: 90.68%) 문장/문서분류 태스크 (단락재순위: 66.3%)

언어모델 단위 별 문맥 반영 워드 벡터

한국 ##어 단 ##어는 ... 구 ##성된 ##다.



한국 ##어 단 ##어는 ... 구 ##성된 ##다.

구글의 언어모델 분석 단위

예문: 한국어 단어는 형태소로 구성된다.

〈ETRI 형태소 기반 언어모델과 구글 언어모델 비교〉



2.6 모델 생성

2. 챗봇 추진 방안

▶ epoch 5, batch size 64 로 설정하여 구글 colab에서 GPU를 사용하여 모델을 생성함

slots output : (?, ?, 6)

Model: "model"

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, None)]	0	
input_mask (InputLayer)	[(None, None)]	0	
segment_ids (InputLayer)	[(None, None)]	0	
KorBertLayer (KorBertLayer)	(None, None, 768)	109693440	input_ids[0][0] input_mask[0][0] segment_ids[0][0]
dense (Dense)	(None, None, 256)	196864	KorBertLayer[0][0]
dense_1 (Dense)	(None, None, 128)	32896	dense[0][0]
time_distributed (TimeDistribut	(None, None, 6)	774	dense_1[0][0]

Total params: 109,923,974

Trainable params: 71,109,254

Non-trainable params: 38,814,720

Input layer ->

KorBERT layer ->

Dense layer 1 ->

Dense layer 2 ->

Time Distributed layer

순으로 신경망을 구성



2.7 모델 테스트

2. 챗봇 추진 방안

▶ 생성된 모델로 문장을 입력해서 테스트를 진행함

Enter your sentence:

나는 과일맛 IPA가 좋아

```
text_arr : ['나는_ 과 일 맛 _ i p a 가_ 좋아_']
```

```
input_ids : [[ 2 752 247 46 2017 9 356 477 374 25 5779 3]]
```

```
input_mask : [[1 1 1 1 1 1 1 1 1 1 1 1]]
```

```
segment_ids : [[0 0 0 0 0 0 0 0 0 0 0 0]]
```

```
inferred_tags : [['0', 'taste', 'taste', '0', '0', 'types', 'types', 'types', '0', '0']]
```

```
slot_score : [[0.999964 0.9977043 0.99870586 0.9999871 0.99999166 0.9999106  
0.99992967 0.9998735 0.9999844 0.9999931 ]]
```

▶ 모델 테스트 결과

- ▶ Slot_score : 사용자가 입력한 말을 토큰으로 분해하고 슬롯으로 태깅한 후, 그 토큰이 태깅된 슬롯일 거라고 예측한 점수



2.8 모델 보완 및 완료

2. 챗봇 추진 방안

▶ Batch size와 drop out을 조절하여 모델을 보완

```
[ '7', '도_', '넘는_', '스', '모', '키', '한_', '바이', '젠', ' ', '추', '천', '해', '쥬', ' ' ]  
[ 'abv', 'abv', 'abv', 'flavor', 'flavor', 'flavor', 'flavor', 'types', 'types', 'types', '0', '0', '0', '0', '0' ]  
[ 0.99997973 0.99998474 0.61759806 0.99980325 0.9998      0.9999037  
  0.999624   0.99997187 0.9999664  0.9999486  0.9999918  0.9999926  
  0.9999902  0.99998915 0.99999   ]
```

문제1. flavor인 '스모키한'이라는 단어를 taste로 잡음

문제2. 문장 틀에 없던 '넘는'이라는 표현을 학습 조건에 따라 abv 혹은 none타입으로 다르게 인식함.

▶ '넘는'을 abv로 인식하면서 '스모키한'을 flavor로 인식하도록 설정값을 변경하며 테스트

batch size 128, dropout이 없는 모델로 수정

03. 챗봇 구현





3.1 서비스 구현 예시

3. 챗봇 서비스 구현

▶ 사용자 입력에 따라 챗봇이 반응하도록 알고리즘 설정 후 flask로 구현

생맥이의 맥주 추천

생맥이
Hi 모두들 안녕
나는 생활 맥주 챗봇, 생맥! 어떤 맥주를 좋아하니?

You 23:14
안녕

생맥이 23:14
헤이~~~ 맥주 한 잔 하실??

You 23:15
응

생맥이 23:15
원하는 맥주에 대해 알려줘~~ 종류는? 도수는? 향은? 맛은? 어떤 게 좋아??~?~

Enter your message... Send

생맥이의 맥주 추천

You 23:15
몰라

생맥이 23:15
혹시 맥주 옵션에 대한 설명이 필요하다면 "설명"을, 아니라면 "x"를 입력해줘

You 23:15
설명

생맥이 23:15
맥주 종류는 '에일', 'IPA', '라거', '바이젠', '흑맥주'가 있어.
에일은 풍부한 향과 진한 색이 특징이야.
IPA는 인디아 페일에일의 준말로, 맛이 강하고 씹쌀한 편이지.
라거는 탄산이 많고 가볍고 청량해.
흑맥주는 색이 까맣고 향미가 진하고.
도수는 3도부터 8도까지 있단다.
향은 '과일'향, '홉'향, '꽃'향, '상큼한' 향, '커피'향, '스모키한' 향 등이 있어.
맛은 '단' 맛, '달지 않은' 맛, '씹쓸한' 맛, '쓰지 않은' 맛, '신' 맛, '상큼한' 맛, '시지 않은' 맛, '과일' 맛, '구수한' 맛 등이 있지.

Enter your message... Send

생맥이의 맥주 추천

생맥이
Hi 모두들 안녕
나는 생활 맥주 챗봇, 생맥! 어떤 맥주를 좋아하니?

You 23:06
흑맥주

생맥이 23:06
흑맥주 말이지?
더 고려할 사항이 있니?

You 23:06
아니

생맥이 23:06
알았어! 널 위한 맥주는 바로!! 오늘밤 스타 우트!!

Enter your message... Send

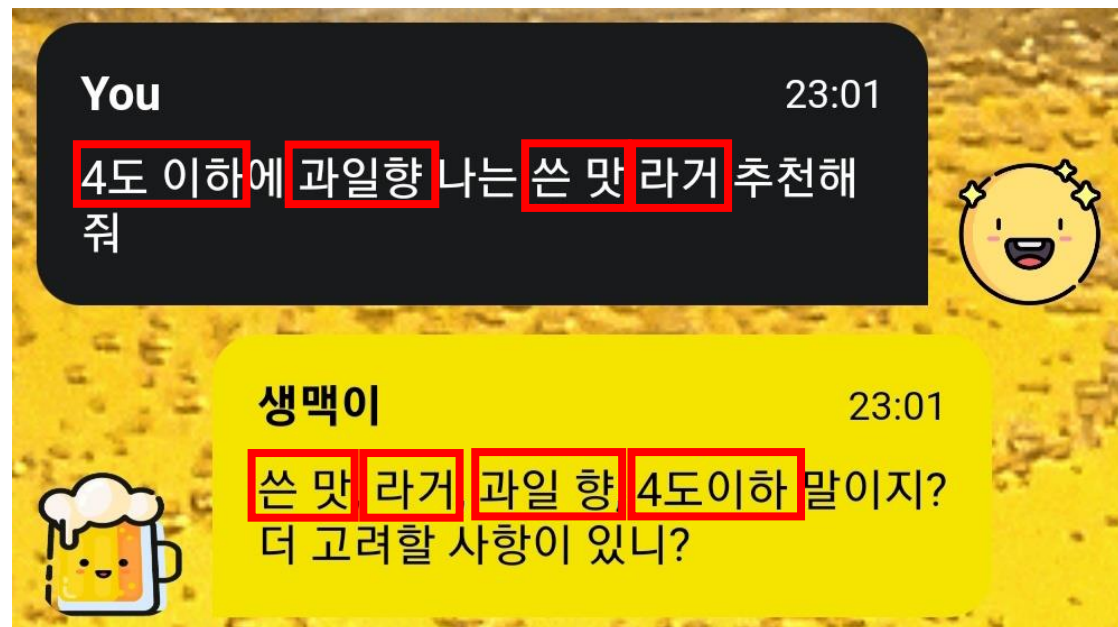


3.2 서비스 구현 알고리즘

3. 챗봇 서비스 구현

- ▶ 사용자의 입력값에서 슬롯을 추출하여 저장한 값을 사용자에게 출력

- ▶ 텍스트에 맞게 태그를 인식
- ▶ 인식한 태그에 맞게 slot_text에 저장
- ▶ 저장한 값을 사용자에게 출력



```
text_arr: ['4', '도_', '이하', '에_', '과', '일', '향', '_', '나는_', '쓴', '맛', '_', '라', '거_', '추', '천',  
input_ids: [[ 2 99 29 3043 13 247 46 480 9 752 4493 2017 9 94  
1206 219 229 59 2377 9 3]]  
input_mask: [[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]]  
segment_ids: [[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]  
inferred_tags: ['abv', 'abv', 'abv', '0', 'flavor', 'flavor', '0', '0', '0', 'taste', '0', '0', 'types', 'types']  
slots_score: [0.9999969, 0.9999987, 0.9999808, 0.9999905, 0.9999962, 0.9999945, 0.9999918, 0.99999845, 0.999999]  
slot_text: {'abv': '4도이하', 'flavor': '과일', 'taste': '쓴', 'types': '라거'}  
app.slot_dict: {'types': ['라거'], 'abv': ['4도', '4도이하', '4도이하'], 'flavor': ['과일'], 'taste': ['쓴']}
```



3.2 서비스 구현 알고리즘

3. 챗봇 서비스 구현

▶ 공통된 토큰값으로 한가지의 맥주를 사용자에게 추천

- ▶ 사용자의 입력 값을 바탕으로 수집한 데이터에서 맥주를 찾음
- ▶ 입력 값에 해당하는 맥주가 있으면 슬롯별로 리스트에 저장
- ▶ 각 슬롯별로 저장된 리스트 중에서 공통된 맥주를 추천



```
rcm_types : ['히트라거', '페스트 비어', '힘내라거', '금빛라거']
```

```
rcm_abv : ['비단밀맥', '호핑몬 페일 에일', '서울.생활.맥주', '페스트 비어', '스마일리 오렌지 윗', '마크퀼쉬', '해문대 페일에일', '설레임',
```

```
rcm_flavor : ['비단밀맥', '호핑몬 페일 에일', '스마일리 오렌지 윗', '해문대 페일에일', '설레임', '제주당 골든에일', '국보IPA', '소나무 IPA',
```

```
rcm_taste : []
```

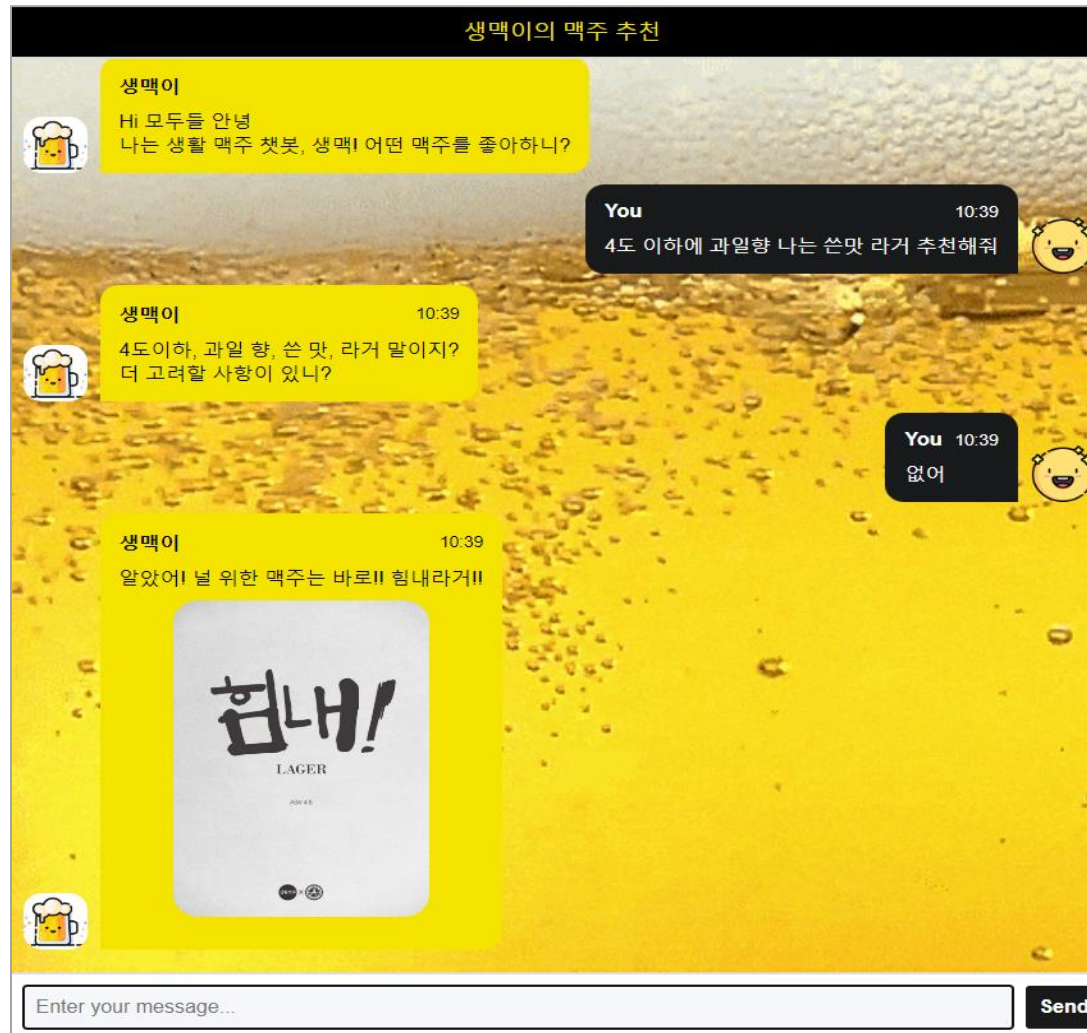
```
최종 추천 제품 : 히트라거
```



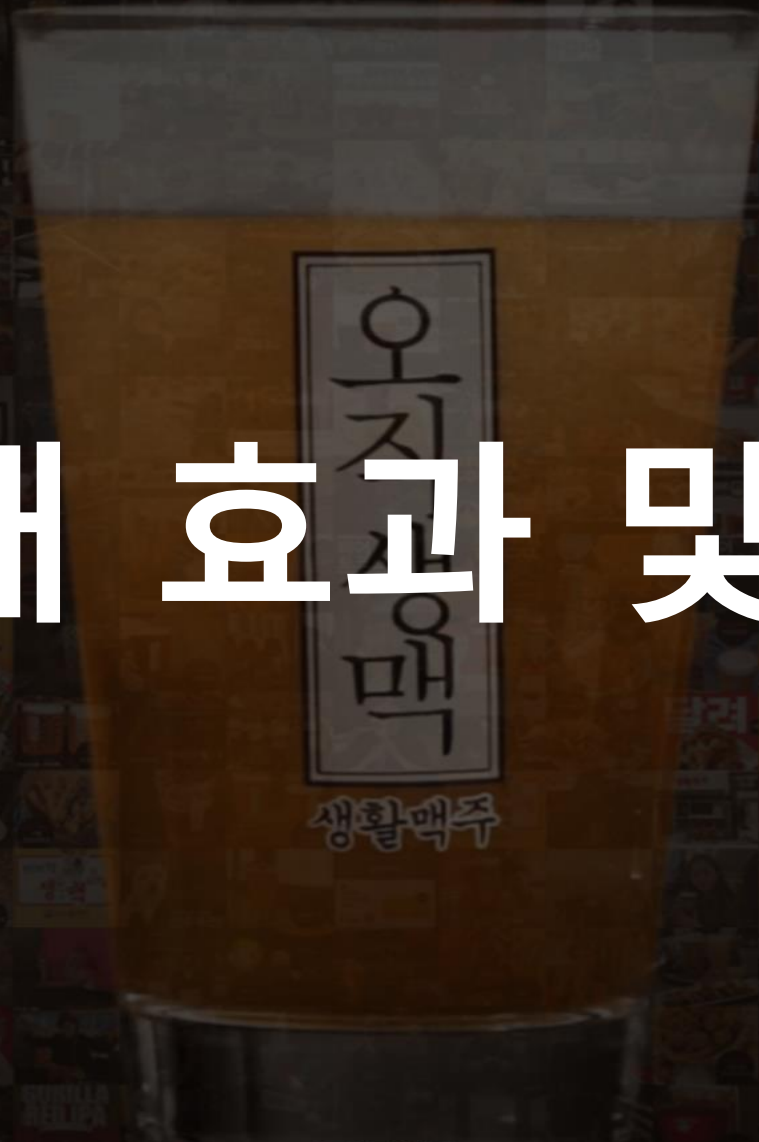

3.3 서비스 시각화

3. 챗봇 서비스 구현

- ▶ CSS, HTML을 사용하여 해당 맥주 이미지 출력 및 챗봇 UI 수정



04. 기대 효과 및 한계점





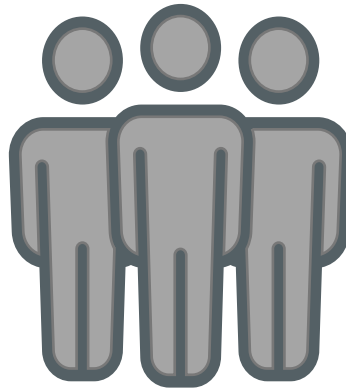
4.1 기대 효과

4. 챗봇 기대 효과

- ▶ 개인화된 추천 챗봇으로 넓은 범위로 발전할 수 있는 서비스 실현 가능성을 확인



개인화된 맥주 추천 방법으로
실현 가능성을 확인



수제 맥주에서 더 확장된 범위의
추천 챗봇으로 발전 가능



수제 맥주를 판매하는 업체에
서비스를 배포하여 주류 산업의
발전 기여



4.2 한계점

4. 챗봇 기대 효과

- ▶ 다양한 문장 틀 확보가 어려움
- ▶ 챗봇의 말투가 비교적 자연스럽지 못함
- ▶ 맛이나 향 등 토큰값을 다양하게 설정하는 데에 어려움
- ▶ 일부 예외처리 외 다양한 상황에서의 예외처리 한계 (정해진 대답에 대한 답변, 불필요한 입력에 대한 답변 등)

05. 소감 및 참고 문헌





5.1 소감

Chapter.5 : 소감 및 참고 문헌



PL/QA

김세리

처음엔 코드 파일과 폴더가 많고 복잡해서 경로 찾기도 쉽지 않았다. 프로젝트 기간 내내 어려운 것 뿐이었지만, 그 이상으로 재미있었고 완성되어가는 모습을 볼 때면 뿌듯했다.

멘토님, 팀원들과 함께 정말 보람찬 시간을 보냈다.

코딩과 협업에 대해 많이 배웠다. 다음에도 챗봇을 만들 기회가 생긴다면 더 잘 할 수 있을 것 같다.



DA

김예지

일상생활에서 자주 접하던 챗봇을 좋은 기회로 멘토님의 도움을 받아 팀프로젝트로 제작하게 되어 기쁩습니다. 팀프로젝트를 진행하면서 혼자서는 할 수 없는 것들을 멘토님과 팀원들의 도움을 받아 할 수 있었고, 내가 어떤 부분이 부족한지를 알 수 있었습니다. 멘토님과 같이 함께한 팀원분들이 열정적하시고 실력도 좋으셔서 많은 걸 배울 수 있는 기회였습니다.



DA

김용관

팀 프로젝트를 시작하기 전 걱정이 앞섰다 실력이 부족하기 때문에 팀원들에게 폐를 끼치지 않을까 하는 걱정과 내가 잘 할 수 있을까 하는 걱정이었다.

하지만 운이 좋게 좋은 팀원들을 만나 무사히 프로젝트를 끝내고 난 후, 코드 한 줄 짜는 것도 힘들던 내가 시간이 오래 걸리더라도 코드를 짤 수 있게 되었고, 그 과정을 즐기게 되어 코딩이 재미있게 느껴졌다.



DA

김다정

데이터 수집부터 모델생성, 챗봇을 구현하기까지 난관이 정말 많았다. 혼자서라면 절대 불가능했을 프로젝트가 팀이었기에 가능했던 것 같다.

개발자로서 부족한 역량들을 이번 프로젝트를 통해 많이 확인했고, 팀원들을 통해 많이 배웠다. 좋은 멘토님과 팀원분들을 만나서 감사했고, 아쉬운 부분도 있지만 좋은 결과물이 나와서 감사하다.



TA/AA

박재형

프로젝트를 진행하기 전에는 스스로 많이 부족하다고 생각했다.

하지만 잘 해내야겠다는 생각으로 열심히 임했고, 완성되어가는 결과물을 보며 성취감과 흥미를 느낄 수 있었다.

부족한 부분들을 멘토님, 팀원들과 함께 채워나가는 과정을 경험하면서, 코딩이나 협력하는 방법에 대해 한 단계 발전한 것 같다.



5.1 참고 문헌

국외논문

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

News

이학수, "전남도, 지역별 수제맥주 육성 관광상품화", 브레이크뉴스, 2022년 06월 29일

송고, "수입맥주 지고 국산 수제맥주 뜬다...3년만에 2.7배 성장", 연합뉴스, 2021년 05월 02일

박현진, "인공지능(AI) 언어모델 'BERT(버트)'는 무엇인가", 인공지능신문, 2019년 01월 03일

인터넷 자료

트랜드모니터, "수제 맥주 인식", 2022년 5월 31일 참고,

<https://trendmonitor.co.kr/tmweb/trend/allTrend/detail.do?bldx=1257&code=0301&trendType=CKOREA#>

엔트의 개발과 일상 블로그, "bert 모델", 2022년 6월 8일 참고,

<https://blog.naver.com/PostView.nhn?blogId=antler07&logNo=222066071573&parentCategoryNo=&categoryNo=44&viewDate=&isShowPopularPosts=true&from=search>

Tmax.ai, "bert", 2022년 6월 14일 참고, <https://tmaxai.github.io/post/BERT/>

WikiDocs, "버트(Bidirectional Encoder Representations from Transformers, BERT)", 2022년 6월 17일 참고, <https://wikidocs.net/115055>

AI PLUS Tech Blog, "Bert", 2022년 6월 24일 참고, <https://blog.est.ai/2019/11/task-oriented-dialog-systems-meet-bert>

Q&A

감사합니다

