(1)

안녕하세요 K-Learning AI 솔루션 발표를 진행하게된

팀장 최지원 입니다.

(2)

**//**본격적인 발표에 앞서 저희 팀을 소개하겠습니다.

팀 이름이 얼리버드인 이유는

학원 입실이 9시까지 이지만

저희 팀원들은 항상 남들보다 30분 일찍 입실하여

하루를 시작했습니다.

이러한 자발적인 모습을 보고

일찍 일어나는 새가 벌레를 잡는다 라는 말 문득 떠올라

얼리버드라는 팀명이 탄생하게 되었습니다.

(3)

**//.**프로젝트 목차는 다음과 같습니다.

서비스 기획,

데이터 탐색 및 전처리,

사용된 모델과 기술,

서비스 제작,

서비스 기대 효과 및 개선점으로 진행되겠습니다.

(4)

**//.**프로젝트에 사용된 기술들은 다음과 같습니다.

(5)

**//.**먼저 서비스 기획에 대해 말씀드리겠습니다.

(6)

**//.**다음은 해외 대학 한국학 현황입니다.

한국학 운영중인 대학은 총 107개국 천여개의 대학으로

일본,중국 ,미국,대만,러시아 순으로 운영되고 있으며,

한국학의 대한 관심도가 점점 증가하고 있습니다.

(7)

**//.**전 세계에서 세종학당을 통해

한국어와 한국 문화에 대해 공부를 할 수 있는데요.

세종학당 수강생이

2007년부터 2021년까지 꾸준히 증가하는 추세를 보였습니다.

**//.**수강생이 폭발적으로 늘어나면서

오프라인 한국어 교육으로는

외국인 학습자의 증가를 감당하기 어려운 상황이라고

김지형 교수님이 인터뷰에서 말씀하셨습니다.

(8)

**//.**그래서 저희팀은 프로젝트의 방향을

외국인들을 위한 한국어 학습 솔루션으로 정하게 되었습니다.

이 서비스를 이용할 학습자의 입장이라고 생각했을때

수준에 맞지 않는 학습을 한다면

큰 어려움을 느끼며 흥미를 잃게 될 것입니다.

**//.**이를 방지하기 위해 레벨 1부터 레벨 3까지

단계별로 학습 솔루션을 제시하여,

학습자가 공부에 흥미를 잃지 않고

한국어 실력이 증진 될 수 있도록 설계했습니다.

(9)

**//.**각 레벨에 맞는 학습 콘텐츠를

구성한 내용은 다음과 같습니다.

레벨 일(1)에서는 BTS, 오징어 게임 등

해외에서 K 문화의 인기가 상승한 것을 이용하여

kpop 가사를 활용한 학습을 설계했습니다.

**//.**레벨 이(2)에서는 좀 더 수준을 높여서

AI-Hub의 일상대화 데이터를 사용하여

학습자에게 일상 대화 형식을 공부할 수 있도록 설계했습니다.

**//.**레벨 삼(3)에서는

한국인들이 영자신문을 이용하여 영어공부를 하는것처럼

외국인들도 이와 같은 방법으로 공부할 수 있지 않을까? 라는 생각이 들어서

다음 뉴스를 이용한 심화 학습을 기획했습니다.

**//.**오른쪽에 나온 표는 프로젝트를 진행하면서 사용한 데이터들 입니다.

(10)

**//.**서비스는

문제 정의,

데이터셋 구축,

모델 개발,

서비스 구현으로 기획되었습니다.

(11)

**//.**데이터 탐색 및 전처리 부분에 대해 말씀드리겠습니다.

(12)

**//.**저희팀이 주로 사용한 데이터는

데이콘에서 열린 문장 유형 분류 AI 경진대회 데이터 입니다.

**//.**유형, 시제, 극성, 확실성 이렇게 크게 4가지로 나눠져 있고

각 레이블별 카테고리들은 화면과 같이 구성되어 있습니다.

(13)

레이블들을 확인한 후 데이터가 어떻게 이루어져있는지 살펴보았는데요.

**//.**히스토그램으로 시각화를 한 결과,

90% 이상의 데이터들의 문장 길이가

100을 넘지 않는 것을 알게 되었으며

countplot을 사용하여

데이터들이 4개의 레이블 안에서 어떻게 분포되어있는지 확인할 수 있었습니다.

**//.**레이블 시각화 결과는 보시다시피

심한 불균형이 있는것으로 확인되었습니다.

또 다른 특징으로는

**//.**데이터는 전부 뉴스 데이터로 이루어져있어서,

맞춤법에 대한 전처리 작업을 따로 진행하지 않았습니다.

(14)

이 데이터가 4가지 레이블로 나누어져 있지만

실제로 들여다보면

**//.**어떤 기준으로 이러한 카테고리들이 붙게 되었는지 의아한 부분이 많았습니다.

한국인인 저희가 봐도 헷갈리는데

학습자 입장에서는 얼마나 혼라스러울까요.

그래서 이러한 혼란을 줄이고자 각 레이블별로

저희가 정한 기준을 두고 정의를 내리는 작업을 진행하게 되었습니다.

(15)

**//.**레이블별 정의를 내리기위해 가장 먼저 어미 분석을 진행하였습니다.

로베르타 스몰 모델을 이용하여 오답노트를 만들고

회의를 통해 정의를 내렸습니다.

**//.**하지만 시제 레이블를 제외한 나머지 레이블에서는 구분 어려워서

다른 데이터 탐색 방법이 필요했습니다.

(16)

**//.**그래서 사용한 2번째 방법은 키위 형태소 분석 입니다.

키위 형태소 분석이란?

코엔엘피와 소엔엘피의 장점만을 가지고 만든 한국어 형태소 분석기 입니다.

한국어 형태소 분석기 라이브러리이며,

koNLPy 라이브러리와 동일하게 미리 한국어 기반의 사전을 정의하여 형태소를 분석하고

Soynlp(소이엔엘피)를 참고해 만든 함수로

말뭉치로부터 새로운 단어를 추출하여

새로운 명사에 적합한 결과들만 추려내는 분석기입니다.

(17)

**//.**키위 형태소 분석을 통해 각 레이블별로 동시출현 네트워크를 시각화하였고

이 시각화를 바탕으로 레이블별 정의를 내렸습니다.

**//.**데이터를 분석한 결과,

영화, 한국, 미국, 중국 등의 단어가 많이 출현한것을 보고

뉴스 데이터의 내용이 많이 편중되어 있다는 것을 알게 되었습니다.

**//.**하지만 이 방법 역시 논리적으로 정의를 내리기 보단 주관적인 판단이 더 강했기에

레이블별 정의를 어떻게 하면 더 논리적으로 정의할 수 있을까 고민이 되었습니다.

동시출현 네트워크란?

나타나는 단어를 모두 표시한 뒤,

두 단어가 같은 문단에 들어가 있다면 둘 사이를 선으로 연결하게 되면서

단어의 네트워크가 만들어지는 것을 말합니다.

동시출현이란?

한 문장, 문단 또는 텍스트 단위에서 같이 출현한 단어를 가리킵니다.

언어학적 의미에서 의미적 근접성을 가리키며, 단어의 연결을 찾는 데 활용됩니다.

동시출현 네트워크란?

이 개념에서 출발한 동시출현 네트워크는 특정 텍스트 단위에서 공동으로 출현한 단어의 집합적 상호 연결을 표현하는 방식이다.

(18)

**//.**데이터 탐색의 논리성을 키우고자 텍스트 전처리를 위한 불용어 목록의 구축이라는 논문을 사용하여 불용어 목록을 구축하고

LDA를 통해 시각화 하였습니다.

**//**LDA란?

문서의 집합으로부터 어떤 토픽이 존재하는지 알아내기 위한 알고리즘입니다.

어떤 토픽과 단어에 대한 사전 정보가 없는 비지도 생성 모델이기 때문에

각 토픽과 관련한 모든 단어를 쉽게 찾을 수 있다는 장점이 있습니다.

왼쪽에 시각화 결과를 보시겠습니다.

각 원과의 거리는 각 토픽들이 서로 얼마나 다른지 보여줍니다.

**//**만약 두 개의 원이 겹친다면, 이 두 개의 토픽은 유사한 토픽이라는 의미입니다.

오른쪽에 막대 그래프는 토픽을 형성하는 키워드를 확인 할 수 있습니다.

데이터 탐색 결과,

추론형, 긍정, 부정과 같은 추상적인 레이블은 시각화로 봤을때

잘 분류하지 못했습니다.

추론형, 긍정, 부정 같은 추상적인 레이블은 시각화로 봤을 때 뚜렷하지 않음

→ 잘 분류하지 못하고 뭉쳐있음

(19)

**//**이렇게 분석한 레이블별 정의는 엑셀 파일로 정리해두었습니다.

**//**어미분석, 문맥 네트워크 분석, LDA로 3차례 데이터 탐색을 한 결과

각각의 분석에서 나왔던 특징들이 비슷하게 발견되었습니다.

(20)

**//**사용된 모델과 기술입니다.

(21)

**//**TTS 기술은 텍스트를 **//** 오디오 데이터로 바꿔주는 기술이며

서비스를 제작할때 받아쓰기에 사용되었습니다.

**//**크롤링은 웹페이지에서 필요한 데이터를 그대로 가져오는 기술이며

다음 뉴스 카테고리를 가져올 때 사용되었습니다.

**//**달리2는 영어 텍스트를 입력하거나 이미지 파일을 삽입하면

자동으로 그림을 생성해주는 기술이며

이미지 생성 문제에서 사용되었습니다.

(22)

**//**다음으로 저희가 사용한 모델 설명이 있겠습니다.

**//**분류 모델은 클루 로베르타 스몰을 사용하였으며

자세한 내용은 다음 슬라이드에서 설명하겠습니다.

**//**마스크 모델은

문장 중간에 빈칸을 만들고

해당 빈칸에 어떤 단어가 적절할지 맞추는 모델이며,

**//**sts 모델은

입력된 문장과 주어진 문장간의 유사도를 볼 때 사용됩니다.

**//**T5 문서 요약 모델은

사용자가 요약한 내용과의 유사도를 측정하는 방법이며,

**//**QA 모델은

문장과 단락의 의미 검색을 위해 설계된 모델입니다.

**//**질문 생성 모델은

BERT 기반으로 문단을 주면

그 문단에서 나올 수 있는 질문들을 자동으로 생성하는 모델이며,

**//**MRC 모델은 BERT 기반으로 사람이 하는 질문과

주어진 문서의 내용을 기계가 이해하고

자동으로 대답할 수 있는 모델입니다.

(23)

**//**로베르타 스몰의 모델을 설명하기 전에

이 모델의 기반이 되는 클루에 대해 먼저 소개하겠습니다.

클루는 한국어 언어 모델의 공정한 평가를 위한 목적으로 만들어진 벤치마크이며

총 8개 종류의 데이터셋으로 구성되었습니다.

**//**이 데이터셋을 만든 연구팀은 baseline model로 Klue-RoBERTa모델을 사용했고

저희팀은 이 모델을 프로젝트에 사용했습니다.

벤치마크란 영상처리에서 MNIST,MS COCO, 자연어처리에서 GLUE, SqUAD, SST와 같이 labeling이 되어 있는 데이터셋을 의미하며 각종 모델들의 성능을 평가할 수 있는 데이터셋(벤치마크)를 의미합니다.

(24)

**//**로베르타 (Robustly Optimized BERT Pretraining Approach) 모델은 크게 3가지 size로 구성됩니다.

기존 BERT와 달리 로베르타는 마스킹이 고정되지 않아서 정확도가 더 올라갈 수 있었습니다.

**//**3가지 사이즈 중에서 로베르타 스몰 모델을 선택한 이유는

사용할수 있는 자원의 한계(google colab)를 고려할때

단시간 내에 다양한 테스크의 구현을 하기 위해서

small 버전을 이용해서 다운스크림 테스크를 구현했습니다.

downstream task 뜻

최종적으로해결하고자 하는 작업

(25)

**//**이제 클루 로베르타 스몰에 대한 코드를 설명하겠습니다.

hugginface에서 pretrain 된 모델을 가져오기 위해서 transformers를 설치합니다.

model은 pytorch를 사용해서 훈련하였습니다.

CFG 사전을 통해 프로그램에 대한 구성 정보를 저장하는데 사용했으며,

seed\_everything 함수를 만들어 모든 시드를 고정하였습니다.

처음에 train과 validation을 train\_test\_split을 사용하여 0.2로 나눴지만

이후에는 성능개선을 위해서 kfold로 대체하였습니다.

CFG가 뭐지?

chat gpt에 물어봄

CFG 사전은 프로그램에 대한 구성 정보를 저장하는 데 사용되는 변수입니다.

코드를 기반으로 CFG는 seed\_everything 함수에 인수로 전달되는 SEED 값을 포함하여 프로그램에서 사용되는 다양한 설정 및 매개변수에 대한 정보를 보유하는 사전일 가능성이 높습니다.

(26)

**//**베이스 모델로 클루 로베르타 모델을 사용하였고 그에 맞추어 **//**토크나이저도 설정하였습니다.

**//**다음으로 레이블 인코딩 작업을 진행하여

카테고리형 데이터를 수치형 데이터로 변환해주었습니다.

이 작업은 머신러닝 모델을 돌려보기 위한 필수 전처리 작업입니다.

train / test / validation 데이터에 라벨링 작업을 수행합니다.

(27)

**//**custom dataset class를 생성하여 타겟 레이블들 설정해 주었습니다.

**//**레이블은 멀티라벨, 멀티 클래스로 구성됩니다.

타입(사실, 추론 , 예측, 대화) // 극성(긍정 미정 부정) // 시제(과거 현재 미래) // 확실성(확실 불확실로)

(28)

**//**hugginface에서 기본적으로 pretrained 된 모델을 가져온 후

타겟 레이블을 출력으로 갖는 layer를 추가하여 **//** CustomModel 클래스를 생성하였습니다.

(29)

**//**label이 4개가 되며 각각 다른 loss가 발생되기 때문에

어떻게 loss를 설정할 것인가를 선택해야 합니다.

**//**저희팀은 대회에서 score 기준인 곱연산을 통해서 loss를 설정했습니다.

이러한 방식을 통해서 각각의 loss가 모두 줄어들게해서

최종 loss가 작아질 수 있도록 하였고

각각 label이 train가능하도록 설정하였습니다.

(30)

**//**마지막으로 만들어진 모델을 train 시키면 기본적인 모델 구축이 완성됩니다.

(31)

**//**구축된 모델의 정확도를 올리기 위해서

저희팀은 첫번째 방법으로 데이터 증식을 사용하였습니다.

데이터 증식 방법을 사용하게 된 이유는

**//**72가지의 카태고리 조합 중

사실-긍정-과거-확실로 편중되어있는것을 볼 수 있습니다.

**//**심지어는 1개의 데이터만 존재하거나

없는 데이터들도 있었습니다.

(32)

**//**따라서 이러한 편향된 데이터를 처리해주기 위해

데이터 증식을 사용하였습니다.

**//**부족한 데이터는

랜덤교체, 랜덤삭제 작업을 통해서 데이터 증식 작업을 진행했으며

불용어 처리를 통해서 성능이 향상될 수 있는지도 탐색했습니다.

(33)

**//**모델 정확도 향상을 위해 사용한 두번째 방법으로

kfold와 앙상블을 사용하였습니다.

5 fold로 쪼갠 후 각각에 대한 model을 train하고

각각의 model 예측결과를

hard voting으로 ensemble 하였습니다.

**//**그후에 데이콘에 모델을 제출하여

베이스 모델과 스코어를 비교해보니

0.72에서 0.74로 정확도가 향상되었습니다.

hard voting

각각의 모델들이 결과를 예측하면 단순하게 가장 많은 표를 얻은 결과를 선택하는 것

ensemble

쉽게 말하면 비슷한 무리들의 집합

즉, 우리는 단일 모델에서 나오는 성능의 결과를 기대하는 것이 아니라,

여러 개의 단일 모델들의 평균치를 내거나,

투표를 해서 다수결에 의한 결정을 하는 등

여러 모델들의 집단 지성을 활용하여 더나은 결과를 도출해 내는 것에 주 목적이다.

kfold와 augmentation을 추가할 경우 너무 시간이 오래 걸리는 관계로

실행가능한 측면을 고려했을때

small 모델로 이후의 작업을 수행하였습니다.

코랩 스탠다드 gpu 기준 train-test-split의 경우

small모델은 30분 안쪽으로 train이 마무리 되었지만

large모델의 경우 하루정도가 소요되었습니다.

(34)

**//**다음은 서비스 제작 입니다.

(35)

**//**지금 줌 채팅으로 링크를 보냈습니다.

엔지록 서버가 터질 수 있기에

기업 관계자분들만 접속하시고

다른 분들은 시연 영상으로 함께 해주시면 감사하겠습니다.

(36)

**//**다음으로 각 단계별 문제 설명이 있겠습니다.

레벨 1에서는 학습자의 흥미를 유발하기 위해

Kpop 가사로 한국어 공부를 기획했는데요

**//**첫번째 문제는

TTS 기술을 사용하여 kpop 가사 일부분을 추출하여 mp3 녹음 파일로 저장후

생성된 음성을 듣고 받아쓰기 하는 형태 입니다.

**//**두번째 문제는 마스크 모델을 사용하여

top\_k로 500개 예측 단어를 출력 후

상위 3개와 하위 1개를 추출하여 객관식으로 주고

노래 가사 중 가장 어울리지 않는 단어를 선택해 달라는 형태 입니다.

**//**세번째 문제는

추출한 가사를 영어로 번역한 후

달리2 기술을 사용하여 이미지를 생성하고

다음 이미지와 맞는 문장을 골라달라는 형태로 구성했습니다.

(37)

레벨 2에서는 살짝 난이도를 높여서 일상 대화를 통한 학습을 기획했습니다.

**//**첫 번째 문제는 문장 유형을 분류하는 모델을 사용하여

앞서 구축한 레이블들을 객관식으로 주고

이 문장에 적합한 레이블을 고르는 문제입니다.

**//**두 번째 문제는 비슷한 문장 선택하기 입니다.

대화 내용에서 5개의 문장을 추출한 후 STS Model을 사용하여

한 문장이 주어졌을 때 그 문장과 가장 비슷한 문장을 선택하는 문제입니다.

(정답은 코사인 유사도를 활용하여 확인함.)

**//**세 번째는 QA 모델을 사용한 위로봇과 대화하기 입니다.

BERT의 문장 임베딩을 이용한 한국어 QA봇을 컨텐츠로 넣었습니다.

유알엘 라이브를 통해 trian\_data를 다운받음.

(38)

레벨 3에서는 마지막 단계인 만큼 뉴스 기사를 통한 한국어 공부를 기획했습니다.

**//**첫 번째 문제는 카테고리 선택하기 이며,

뉴스의 내용을 보고 적절한 분류를 선택하는 문제입니다.

크롤링을 이용하여 실시간 다음 홈페이지에 있는 뉴스를 끌고와

매번 최신 뉴스로 학습할 수 있다는 장점이 있습니다.

**//**두 번째 문제는 문장 요약하기 입니다.

학습자가 고른 뉴스를 요약하라는 문제를 주고

T5 model을 사용하여 얻은 요약 문장을 정답으로 지정했습니다.

채점 방식은 사용자가 요약한 문장과 모델로 얻은 요약 문장을

STS model을 사용하여 코사인 유사도를 측정하고

일정값을 넘기면 정답 처리하는 방식으로 설계되었습니다.

**//**세 번째 문제 역시 문장 유형을 선택하는 문제이며,

뉴스 본문에 있는 문단을 랜덤하게 추출하여 T5 model로 요약한 뒤에

요약문을 가지고 문장 유형을 물어보는 문제입니다.

**//**네 번째 문제는 큐앤에이 문제입니다.

kobert-QuestionGeneration을 사용하여 뉴스 문단의 내용에 맞는 질문을 생성하고

학습자가 질문에 대한 답을 작성하면

klue-bert-base-aihub-mrc 모델을 사용하여 생성된 답을 학습자에게 제시되도록 설계되었습니다.

첫번째 문제

[문제 생성] 연합 뉴스 토픽 분류 라벨링 된 데이터 활용하여 카테고리 분류

다음 뉴스 실시간 크롤링 -> 카테고리 라벨링

[채점] 뉴스를 읽고 해당 뉴스 카테고리 선택 & 라벨링된 카테고리로 정답 확인

두번째 문제

label 되지 않은 데이터셋을 임의로 mask하여 train해서 생성에 이용

T5 Model을 사용하여 요약 -> 정답

**Text-to-Text Transfer Transformer-eenzeenee/t5-small-korean-summarization**

사용자가 요약한 내용과 유사도를 측정 -> STS 문장 단위 임베딩한 후 코사인 유사도 측정하여 일정값 (0.5)을 넘는 경우 정답으로 인정

세번째 문제

앞에서 구축한 문장 유형 분류 모델을 활용하여 문장이 하나 주어졌을 때 그 문장의 시제, 극성, 확실성 선택

이 문제는 문장유형분류AI경진대회 데이터 셋을 이용

klue/roberta-small모델에 downstream task를 fine tuning하여 만들어 보았습니다.

긍정, 부정, 미정의 경우 대회 데이터로는 부족하다고 판단하여

NSMC 데이터를 기존 모델에 다시 Training 한 후

score에 따라 긍정 > 미정( 중간단계) > 부정 순서로 라벨링 하였습니다.

단순히 문장의 종결어미가 아닌 실제 문장의 전달하는 내용을 기준으로

긍정/미정/부정을 판단하면 됩니다.

네번째 문제

[질문생성] Sehong/kobert-QuestionGeneration

문단이 주어지면 그 문단에서 나올 수 있는 질문들을 자동으로 생성

[질문 답변] bespin-global/klue-bert-base-aihub-mrc

사람이 자연어로 하는 질문과 주어진 대상 문서의 내용을 기계가 이해하고

자동으로 대답할 수 있는 기술

뉴스 내용에 대한 질문이 생성되면 정답을 사용자가 입력

-> 질문 답변 기술로 해당 질문에 대한 답변을 알려줌.

(39)

**//**이렇게 구현한 서비스를

외국인 친구에게 사용을 부탁하고

관련 피드백을 받았습니다.

뉴스 문제는 좀 어려웠지만

레벨 별로 문제 난이도가 적당했으며,

한국어를 배우는 입장에서 이런 사이트가 많이졌으면 좋겠다고 후기를 남겨주었습니다.

(40)

**//**서비스 기대효과 및 개선점 입니다.

(41)

**//**서비스 기대효과로,

첫번째 실생활 데이터들로 인해 회화학습에 도움이 될 수 있습니다.

두번째는 노래가사, 실시간 뉴스로 한국의 문화를 다양하게 체험할 수 있습니다.

마지막으로 수준에 맞는 단계별 학습을 통해 한국어 실력 증진 효과를 기대할 수 있습니다.

(42)

**//**프로젝트 개선점입니다.

**//**기술 부분에서는 모델 정확도 개선과 최적화 작업이 필요합니다.

그리고 자연어의 꽃은 챗봇이라고 말할 수 있는데요.

이 서비스에서 저희팀은 기초적인 QA봇을 사용했기때문에

챗봇 구현의 아쉬움이 남았습니다.

**//**서비스 부분에서는 학습자에게 좀 더 다양한 컨텐츠가 필요합니다.

(43)

**//**한달간 프로젝트를 같이 진행한 팀원들의 소감 입니다.

(44)

**//**사용한 모델, 이미지, 자료 등의 출처를 남기면서

(45)

**//**K-Learning AI 솔루션 발표를 마치겠습니다.

감사합니다.