

KUBIG-DL분반
NLP 프로젝트
최종 결과 발표

14기 김수경, 김유민, 임혜리, 채윤병
15기 김지후



목차

01

프로젝트 소개

Overview

대회 소개 01

데이터 02

02

주요 과정

Main process

전처리 03

모델링 04

03

최종 결과 및 마무리

Result & Summary

결과 및 맺음말 05

01 Overview

프로젝트 소개

대회 소개	01
데이터	02

한국어 문장 관계 분류 경진대회

월간데이콘18 | KLUE | 자연어 | 한국어 | ACCURACY

₩ 상금 : 100만원 + α

🕒 2022.01.28 ~ 2022.02.28 17:59

[+ Google Calendar](#)

👤 768명 📅 D-11

Dacon 한국어 문장 관계 분류 경진 대회

📌 대회 개요

두 문장의 관계를 분류 - NLI(Natural Language Inference)

카테고리 - 참(Entailment), 거짓(Contradiction), 중립(Neutral)

평가지표 : 정확도(Accuracy)

상위 10개 팀 입상

2월 28일 종료

premise ▼	hypothesis ▼	label
씨름은 상고시대로부터 전해져 내려...	씨름의 여자들의 놀이이다.	contradiction
삼성은 자작극을 벌인 2명에게 형사...	자작극을 벌인 이는 3명이다.	contradiction
이를 위해 예측적 범죄예방 시스템...	예측적 범죄예방 시스템 구축하고 고도화...	entailment
광주광역시가 재개발 정비사업 원주...	원주민들은 종합대책에 만족했다.	neutral
진정 소비자와 직원들에게 사랑 받...	이런 상황에서 책임 있는 모습을 보여주...	neutral
이번 증설로 코오롱인더스트리는 기...	코오롱 인더스트리는 총 9만 3800톤의 생...	entailment
자신뿐만 아니라 남을 돕고자 하는 ...	모든 청년은 꿈과 열정을 가지고 있다.	neutral

데이터 설명

Premise 문장과 hypothesis 문장, label로 이루어진 텍스트 데이터
데이터 출처 : Klue
train : 약 25000개, test : 약 1700개

Premise와 hypothesis 문장의 관계를 분류

02

Main process

주요 과정

전처리	03
모델링	04

전처리 방법

1.데이터 추가

KLUE에서 데이터 추가
(10%)

2.텍스트 전처리

한글과 숫자를 제외한 특
수문자 제거(마침표, 따옴
표 등)

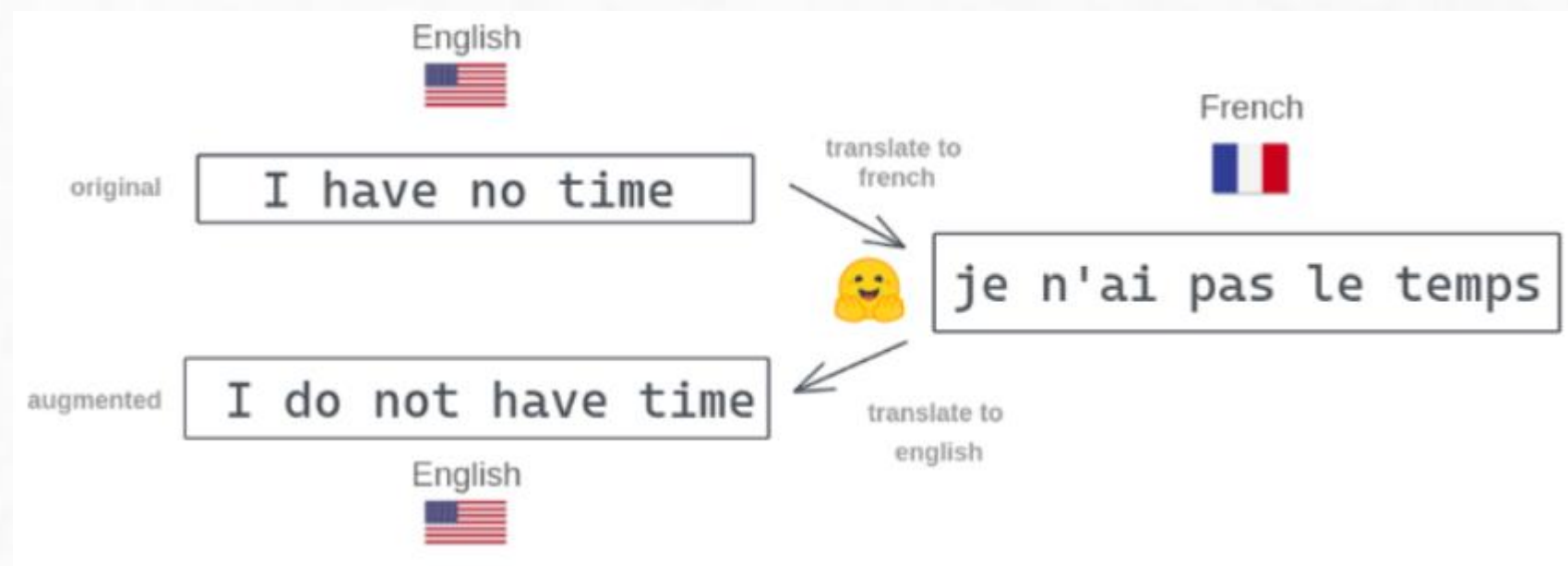
3.번역 전처리 (Back translation)

한글 > 영어 > 한글

◆ 이 외에도 형태소 분석 전처리도 해보았으나 효과적이지 않았음

Back Translation (KOR→ENG→KOR)

- Back translation은 monolingual한 데이터셋을 번역기를 통해 인위적으로 augmentation하는 방법
- Label을 유지한 채로 원본 데이터를 타 언어로 번역한 뒤, 다시 원래의 언어로 재 번역하는 과정을 통해 다양성을 확보
- 이때 학습된 번역기 구축 작업이 필요하지만 단순 label 분류문제를 위해선 새로운 번역 모델을 만드는 것은 비효율적이라 판단
따라서 기존에 구축되어 있는 PAPAGO, Googletrans 번역기 활용



번역 전처리 예시

Back Translation (KOR→ENG→KOR)

< 단계 설명 >

1. 번역기를 통해 train data의 text(한국어)를 영어로 번역
2. 영어로 번역된 text를 다시 한국어로 번역
3. 앞선 두 번의 번역을 거쳐 생성된 새로운 문장을 train data에 추가하여 Data augmentation 진행

< 기대 효과 >

"Back translation을 통해 유사한 의미를 가진 동의어로 단어가 대체되기도하고, 순서가 바뀌기도 하며 오히려 조금은 부자연스러운 문장들이 형성됩니다. 이렇게 생성된 어색한 문장은 학습 시 강한 train signal을 발생시키기 때문에 과적합을 줄이거나 예측력을 높이는 것에 도움이 됩니다."



참고논문 : Edunov et al. "Understanding Back-Translation at Scale" EMNLP. 2018

Back Translation (KOR→ENG→KOR)

```
[ ] def kor_to_trans(text_data, trans_lang, start_index, final_index):

    target_present = EC.presence_of_element_located((By.XPATH, '//*[@id="txtTarget"]'))

    for i in tqdm(range(start_index, final_index)):

        if (i!=0)&(i%33==0):
            driver.implicitly_wait(2)
            print('{}th : '.format(i), backtrans)
            np.save(data_path+'kor_to_eng_train_{}_{}.npy'.format(start_index, final_index), trans_list)

        try:
            driver.get('https://papago.naver.com/?sk=cost&src='+trans_lang+'&src='+text_data[i])
            time.sleep(1.5)
            element=WebDriverWait(driver, 10).until(target_present)
            time.sleep(0.1)
            backtrans = element.text

            if (backtrans=='')|(backtrans==' '):
                element=WebDriverWait(driver, 10).until(target_present)
                backtrans = element.text
                trans_list.append(backtrans)
            else:
                trans_list.append(backtrans)

        except:
            trans_list.append('')
```

크롤링(Papago)

But, 오번역되는 경우는 여러가지 케이스가 존재

<재번역 대상>

번역이 이뤄지지 않아 null값을 갖는 경우 (시간
제한으로 인해 약 500여개 문장 번역 x)
번역된 문장이 기존 문장의 길이에 대한 비율이 0.
2이하인 경우

고유명사 영어 -> 한글로 수정

03. 전처리

KUBIG

KOR->ENG 예시

1 to 25 of 20000 entries Filter ?					
index	premise	hypothesis	eng_premise	eng_hypothesis	label
0	씨름은 상고시대로부터 전해져 내려오는 남자들의 대표적인 놀이로서, 소년이나 장정들이 넓고 평평한 백사장이나 마당에서 모여 서로 힘과 술기를 겨루는 것이다.	씨름의 여자들의 놀이이다.	Ssireum is a representative game of men that has been handed down since the ancient times, and boys and generals gather in a wide and flat white sandy beach or yard to compete for strength and wisdom.	It's a play of women in Ssireum.	contradiction
1	삼성엔 자작극을 벌인 2명에게 형사 고소 등의 법적 대응을 검토 중이라고 하였으나, 중국 내에서의 여론은 자작극이라는 증거가 충분함에도 불구하고 좋지 않다.	자작극을 벌인 이는 3명이다.	Samsung said it was considering legal action, including criminal charges, against two people who made a self-made play, but public opinion in China is not good despite sufficient evidence that it is a self-made play.	There were three people who made their own plays.	contradiction
2	이를 위해 예측적 범죄예방 시스템을 구축하고 고도화한다.	예측적 범죄예방 시스템 구축하고 고도화하는 것은 목적이 있기 때문이다.	To this end, a predictive crime prevention system is established and advanced.	This is because the purpose is to establish and upgrade a predictive crime prevention system.	entailment
3	광주광역시와 재개발 정비사업 원주민들에 대한 종합대책을 마련하는 등 원주민 보호에 적극 나섰다.	원주민들은 종합대책에 만족했다.	Gwangju Metropolitan City has actively started to protect natives by preparing comprehensive measures for the redevelopment and maintenance project natives.	The natives were satisfied with the comprehensive measures.	neutral
4	전정 소비자와 직원들에게 사랑 받는 기업으로 오래 지속되고 싶으면, 이런 상황에서는 책임 있는 모습을 보여주는 것이 필요하다.	이런 상황에서 책임 있는 모습을 보여 주는 기업은 아주 드물다.	If you really want to last a long time as a company loved by consumers and employees, it is necessary to show responsibility in this situation.	In this situation, very few companies show responsibility.	neutral
5	이번 증설로 코오롱인더스트리는 기존 생산량 7만 7000톤에서 1만6000톤이 늘어나 총 9만 3800톤의 생산 능력을 확보하게 됐다.	코오롱 인더스트리는 총 9만 3800톤의 생산 능력을 확보했다.	With the expansion, Kolon Industries has increased its production capacity by 16,800 tons from 77,000 tons to a total of 93,800 tons.	Kolon Industries has secured a total production capacity of 93,800 tons.	entailment
6	자신뿐만 아니라 남을 돕고자 하는 청년의 꿈과 열정에 모두가 주목하고 있다.	모든 청년은 꿈과 열정을 가지고 있다.	Everyone is paying attention not only to themselves but also to the youth's dreams and passion to help others.	All young people have dreams and passion.	neutral
7	시대상황을 고려하는 현명한 시청태도가 요구된다.	시청태도에 특별한 주의점은 없다.	A wise viewing attitude is required to consider the situation of the times.	There is no special attention to viewing attitude.	contradiction
8	사진과 차이없는 아기자기한 실내소품들과 분위기가 멋졌습니다.	아기자기한 실내소품들은 사진에서 본 것과 차이가 있었습니다.	The cute indoor props and atmosphere without a difference from the picture were wonderful.	The cute indoor props were different from what I saw in the picture.	contradiction
9	빠른 답장과 간편한 체크인, 깨끗한 집 좋았어요	체크인이 복잡했어요.	Quick reply, simple check-in, and clean house were good.	Check-in was complicated.	contradiction
10	대부분 도보로 이동하기 충분하다는 점이 매력적이었어요.	대부분 걸어서 갈 수 있어요.	It was attractive that most of them were enough to walk.	Most of the time, you can walk.	entailment
11	오후에는 소흘읍민의 멋진 재를 엿볼 수 있는 조수들 레질 버스킹 공연이 연이어 진행된다.	조수들레질 버스킹 공연에 많은 군민들이 참석했다.	In the afternoon, there will be a series of busking performances on the lake dulle-gil, where you can see the wonderful talents of Soheul-eup residents.	Many county residents attended the busking performance on the lake dulle-gil.	neutral
12	영화 시작부터 끝까지 긴장감을 놓을 수가 없네요.	영화 시작부터 긴장감이 함께하네요.	I can't relax from the beginning to the end of the movie.	There's tension from the start of the movie.	entailment

03. 전처리

KUBIG

ENG->KOR 예시

1 to 25 of 20000 entries Filter ?					
index	premise	hypothesis	kor_premise	kor_hypothesis	label
0	씨름은 상고시대로부터 전해져 내려오는 남자들의 대표적인 놀이로 서, 소년이나 장정들이 넓고 평평한 벼사자거나 마당에서 모여 서로 힘과 술기를 겨루는 것이다.	씨름의 여자들의 놀이이다.	씨름은 예로부터 전해 내려오는 남자들의 대표적인 놀이로 넓고 평평한 벼사자거나 마당에 소년들과 장군들이 모여 힘과 지혜를 겨룬다.	씨름에 나오는 여자들의 놀이입니다.	contradiction
1	상설을 자작극을 벌인 2명에게 형사 고소 등의 법적 대응을 검토 중 이라고 하였으나, 중국 내에서의 여론은 자작극이라는 증거가 충분 함에도 불구하고 좋지 않다	자작극을 벌인 이는 3명이다.	상설을 자작극을 벌인 2명에 대해 형사고발 등 법적 대응을 검토하고 있다고 밝혔지만 자작극이라는 충분한 증거에도 불구하고 중국 내 여론이 좋지 않다	그런 자신의 연극을 만든 세 사람이 있었다.	contradiction
2	이를 위해 예측적 범죄예방 시스템을 구축하고 고도화한다	예측적 범죄예방 시스템 구축하고 고도화하는 것 은 목적이 있기 때문이다.	이를 위해 예측형 범죄예방시스템이 구축되고 고도화된다	예측형 범죄예방 시스템 구축과 고도화가 목 적이기 때문이다.	entailment
3	경주광역시가 재개발 정비사업 원주민들에 대한 종합대책을 마련하 는 등 원주민 보호에 적극 나섰다	원주민들을 종합대책에 만족했다.	경주광역시가 재개발 정비사업 원주민 종합대책을 마련하는 등 원주 민 보호에 적극 나섰다	그 원주민들은 포괄적인 조치에 만족했습니 다	neutral
4	전장 소비자와 직원들에게 사랑 받는 기업으로 오래 지속하고 싶으 면, 이런 상황에서는 책임 있는 모습을 보여주는 것이 필요하다.	이런 상황에서 책임 있는 모습을 보여주는 기업은 아주 드물다.	소비자와 직원들이 사랑을 받는 기업으로 장한 오래 지속하고 싶다 면 이런 상황에서 책임감을 보여줄 필요가 있다.	이런 상황에서 책임감을 드러내는 기업은 극 소수다.	neutral
5	이번 공설부 고모골엔디스브리는 기존 생산량 7만7000주에서 1만 6800주가 늘어나 총 9만 3800주의 생산 능력을 확보하게 된다	고모골 엔디스브리는 총 9만 3800주의 생산 능 력 을 확보했다	고모골엔디스브리는 이번 공설부 생산능력이 7만7000에서 총 9만 3800으로 1만6800이 늘었다	고모골엔디스브리는 총 9만3800의 규모의 생 산능력을 확보했다	entailment
6	치실뿐만 아니라 입을 돌고자 하는 성인의 골치 신경에 모두가 주목 하고 있다.	모든 성인은 골치 신경을 가지고 있다.	모두가 치실뿐만 아니라 성소인의 골치 입을 돌고자 하는 신경에 관 심을 갖고 있다고 한다.	모든 젊은이들은 골치 신경이 있습니다.	neutral
7	시대상황을 고려하는 선명한 사실확도가 요구된다	시대태도에 특별한 주의점은 없다	시대적 상황을 고려하는 선명한 관련 지체가 요구된다	모든 태도에 특별한 관심이 있는 것은 아니 다.	contradiction
8	사전과 차이없는 아기자기한 실내소품들과 분위기가 멋졌습니다.	이끼지기한 실내소품들은 사진에서 본 것과 차이 가 있었습니디	사전과 차이가 없는 아기자기한 실내 소품과 분위기가 멋졌다.	귀여운 실내 소품들이 사진에서 본 것과 같았 습니디	contradiction
9	빠른 입장과 간편한 체크인, 깨끗한 집 좋았어요	체크인이 복잡했어요.	빠른 입장과 간편한 체크인, 그리고 깔끔한 집이 좋았다	체크인을 복잡했다	contradiction
10	대부분 도로로 이동하기 불편하다는 점이 매력적이었어요.	대부분 골에서 살 수 있어요.	대부분이 골을 수 있을 정도로 넉넉한 것이 매력적이었다.	대부분은 골을 수 있습니다.	entailment
11	오후에는 소풍장면의 멋진 커피를 맛볼 수 있는 포수블레쉴 베스팅 공 연이 연이어 진행된다.	포수블레쉴 베스팅 공연에 많은 관객들이 참석했 다.	오후에는 소풍장 주민들의 멋진 커피를 볼 수 있는 포수블레쉴 베스팅 공연이 이어진다.	포수블레쉴 베스팅 공연에는 많은 관객이 참 석했다.	neutral
12	영화 시작부터 끝까지 긴장감을 놓을 수가 없네요.	영화 시작부터 긴장감이 완결하네요.	저는 영화의 시작부터 끝까지 긴장을 둘 수가 없어요.	영화 시작부터 긴장감이 감동어요.	entailment
13	아글 벅의 자살 정보가 영국에 도착한 것은 7월 16일이였다.	아직 아글 벅의 자살 정보가 영국에 도착하지 않 았다.	제이콥 벅의 자살 정보가 영국에 도착한 것은 7월 16일이였다.	제이콥 벅의 자살 정보는 아직 영국에 도착하 지 않았다.	contradiction
14	가족이 없거나 야간시간대 돌봄을 제공하기 위해 하루 중 단시간 동안 수시로 방문하는 24시간 순회돌봄서비스 도입 등이 그 예다.	24시간 순회돌봄서비스는 가족이 없거나 야간시 간대 돌봄을 제공하기 위해 하루 중 장시간 동안 한번 방문한다.	가족이 없거나 낮에 단기간 자주 방문해 야간 돌봄을 제공하는 24시 간 투여 케어 서비스 도입 등이 대표적이다.	24시간 광복 케어 서비스는 하루 1회 장시간 방문해 가족이 없는 야간에도 케어를 제공한 다.	contradiction

모델링 방법

KoELECTRA, Roberta base, Roberta Large 등 다양한 사전 학습 모델 사용

모델링 목적 : 사전 학습 모델, 학습 조건(epoch, dropout, optimizer) 등을 변화 시켜가며 최선의 모델을 찾는 것

모델링 패키지 : pytorch, tensorflow

Model	koELECTRA-base	KLUE-RoBERTa-base	KLUE-RoBERTa-large
NLI acc	85.63	84.83	89.17



KLUE RoBERTa 모델의 경우 2021년 5월에 발표된 RoBERTa 기반 한국어 자연어 처리 모델
KLUE 자체 벤치마크 점수 기준 기계독해 포함 대부분의 자연어 처리 테스트에서 우수한 성능 보임



각 모델의 사전
학습 시의 성능

모델 앙상블

1. Hard voting

각각의 모델들이 결과를 예측하면 단순히 가장 많은 표를 얻은 결과를 선택

2. Soft voting

각 class별로 모델들이 예측한 probability를 합산하여 가장 높은 class를 선택

3. Seed ensemble

train-test-split 과정에서 seed number 다르게 지정하여 훈련 통해 나온 모델 기하평균

4. fold ensemble

교차검증(cross-validation) 훈련을 통해 나온 모델 기하평균

◆ 최종 모델은 Roberta-Large 기반의 Fold ensemble + KoELECTRA
모델 예측 앙상블

모델 앙상블

학습 1

해석 데이터를 추가하지 않은
데이터로 학습(RoBERTa-Large)

fold1
fold2
fold3
fold4
fold5

학습 2

해석 데이터를 추가한
데이터로 학습(RoBERTa-Large)

fold1
fold2
fold3
fold4
fold5

학습 3

해석 데이터를 기존 train data와
변경(RoBERTa-Large)

↔
학습 3,4는 fold를 나눠서
학습하지 않음

학습 4

해석 데이터를 추가하지 않은
데이터로 학습(KoELECTRA)

fold학습 모델을 모두 앙상블 했을 때
어느 한 모델이 모델 전체의 성능을
감소하는 경우 발생 -> 일부는 제외
최종적으로 빨간색 글씨의 8개 모델
예측 앙상블

03








Result & Summary

최종 결과 및 마무리

결과 및 맺음말 05

최종 결과

상위 10개팀 입상 예정 -> 최종 순위는 6위!
코드 검증만 남은 상황

● WINNER ● 1% ● 4% ● 10%			
#	팀	팀 멤버	최종점수
6	말복딱복		0.89555
1	가온		0.93097
2	snoop2head		0.89975
3	Maximalizm		0.89915
4	휘오		0.89735
5	Lee		0.89615
6	말복딱복		0.89555

느낀점

- 비교적 간단한 task임에도(문장 관계 분류) 시도해 볼만한 것들이 많았음.
- 자연어 처리의 전체적인 과정(전처리, 토큰나이징, 모델링)을 해볼 수 있어서 좋았다.
- 자연어 처리 모델링에서는 역시 pre-trained된 모델이 강력하다..!
- 전처리 방법, 모델 앙상블을 통해 많은 시도를 해보았고 이러한 시도들이 경험적으로 도움이 되었다.

KUBIG-DL(NLP)

THANK YOU

감사합니다.