SUNGKYUNKWAN

JEJURICA :

제주도 방언 음성을 영어 음성으로

CAPSTONE 김윤서 | 김조현 | 키미야

CONTENTS

SUNGKYUNKWAN

SUNGKYUNKWAN

01

문제 정의
Problem definition

- 문제 상황
- 문제 해결 방안

Chapter.01

문제 정의

Problem definition

제주도 방언 번역기의 필요성 + 글로벌화

제주도 사투리 보존 문제

- 2011년 12월 유네스코가 지정하는 ‘소멸위기 언어’ 5단계 중 4단계인 ‘아주 심각한 위기에 처한 언어’에 등재
- 실제 제주도민 10명 중 3~4명은 노인들이 사용하는 제주사투리 이해 못함

한국 문화의 글로벌화

- 제주도로 방문하는 외국인의 수 매년 증가
- 영화, 드라마 등 한류 콘텐츠의 전파에도 방언의 영어 번역은 필수

실질적인 문제 해결 방안 부재

- 실용화된 제주도 방언 번역기 부재
- 대부분의 번역기는 표준어 기준

Chapter.01

문제 정의 Problem definition

제주도 방언 번역기의 필요성 + 글로벌화

제주도 방언 음성을 영어 음성으로 번역하여 생성하는 서비스를 제공하자!

제주도 사투리 보존 문제

- 2011년 12월 유네스코가 지정하는 '소멸위기 언어' 5단계 중 4단계인 '아주 심각한 위기에 처한 언어'에 등재
- 실제 제주도민 10명 중 3~4명은 노인들이 사용하는 제주사투리 이해 못함

한국 문화의 글로벌화

- 제주도로 방문하는 외국인의 수 매년 증가
- 영화, 드라마 등 한류 콘텐츠의 전파에도 방언의 영어 번역은 필수

실질적인 문제 해결 방안 부재

- 실용화된 제주도 방언 번역기 부재
- 대부분의 번역기는 표준어 기준

SUNGKYUNKWAN

02

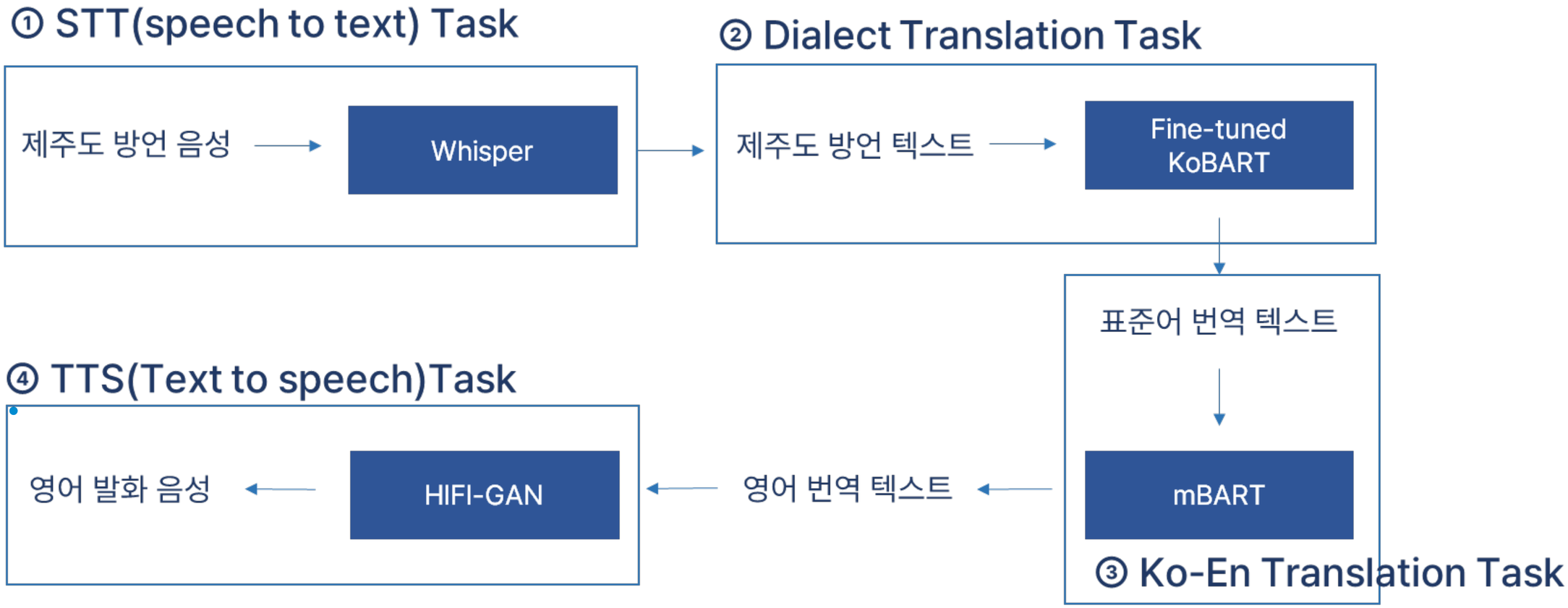
프로젝트 개요
Project flow

- 프로젝트 진행 방향
- 사용한 모델과 그 이유

Chapter.02

프로젝트 개요

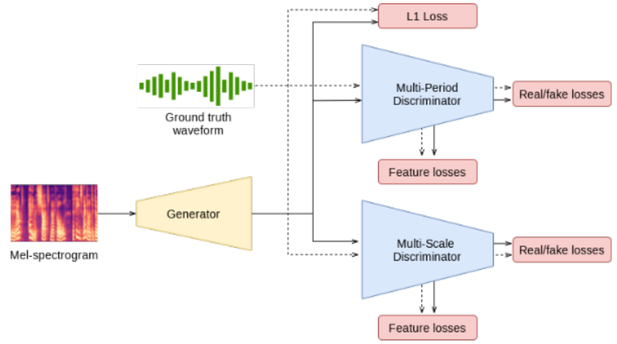
| Project Flow



Chapter.02

사용한 모델 Model

OpenAI Whisper



음성인식 모델 Whisper

한국어 약 8000시간 사전훈련.
제주도 방언 음성 STT성능 네이버 클라바와 비슷(fine-tuned X)
인코더-디코더 트랜스포머 구조

번역 모델 BART(KoBART, mBART)

인코더 디코더를 둘다 가진 구조로
번역성능에 탁월.
한국어로 사전학습 되어
한글 테스크에 적합

TTS 모델 HIFI-GAN

1개의 generator와 2개의 discriinator
로 mel-gan보다 음성 생성에
우수한 성능을 보임.
LJ speech 24시간 데이터로 사전학습

SUNGKYUNKWAN

03

데이터 수집 및 전처리
Data collect

- 데이터 수집 source
- 총 사용한 데이터

Chapter.03

데이터 수집 Data collect

Whisper

성능평가를 위한 데이터

AI HUB 중노년층 방언 음성데이터

(.wav) 제주도

1000개 파일 사용

mBART

성능평가를 위한 데이터

AI HUB 발화유형(문어/구어/채팅)별 기계번역 병렬 말뭉치 중 구어 데이터

500개 파일 사용

KoBART

<데이터 표>

AI HUB	JIT Dataset	제주도 방언 사전 크롤링	총계
84,385	114,626	33,646	232,657

AI Hub

Fine-tuning을 위한 데이터

총 *232,657* 개의 (제주도 - 표준어)쌍 데이터 수집

SUNGKYUNKWAN

04

모델링
Modeling

- train/test 분리
- 하이퍼 파라미터 및 fine-tuning 프로세스

Chapter.04

모델링

Whisper , mBART, Hifi -GAN 모델

- mbart의 경우, 다국어 모델은 불완전한 번역이 있었음

-> 한국어 코퍼스로 추가로 finr-tuning 한 모델 사용

- Whisper의 경우, fine-tuning 진행.

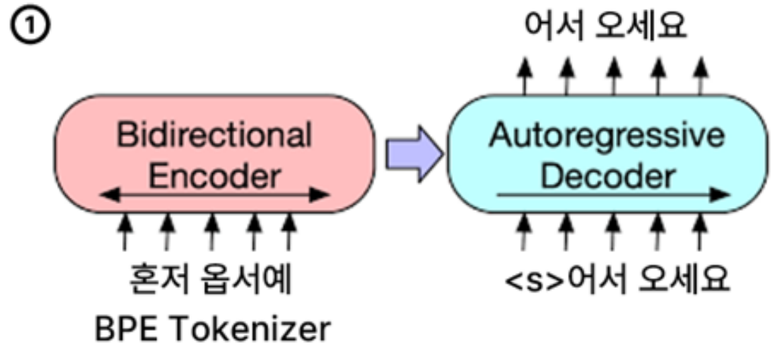
-> 모델이 무거워 시간 소요 & 오류 해결 실패(반복된 문구)

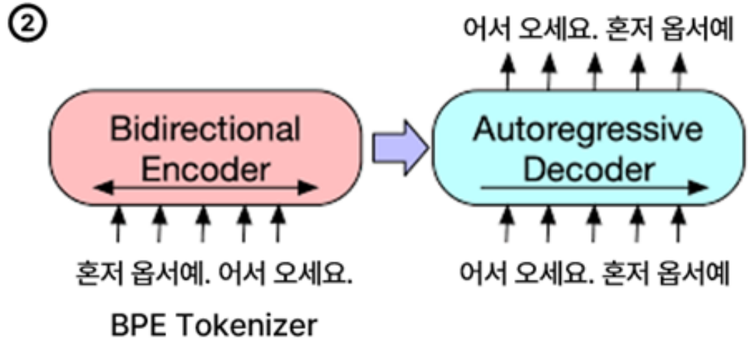
- HiFi -gan 모델 + tactrone 결합 사용

-> 허깅페이스 제공 모델 사용

Chapter.04

모델링 KoBART fine-tuning





인코더에 제주 방언 텍스트

디코더에 표준어 텍스트

데이터를 *src+target, target+src* 와 같은 형태로 구성하여 양방향으로 학습이 될 수 있도록 구성

방법 A : ① 방식으로 full epoch

방법 B : ② 방식으로 1/3 epoch, 가중치 freeze 후 ① 방식으로 나머지 epoch

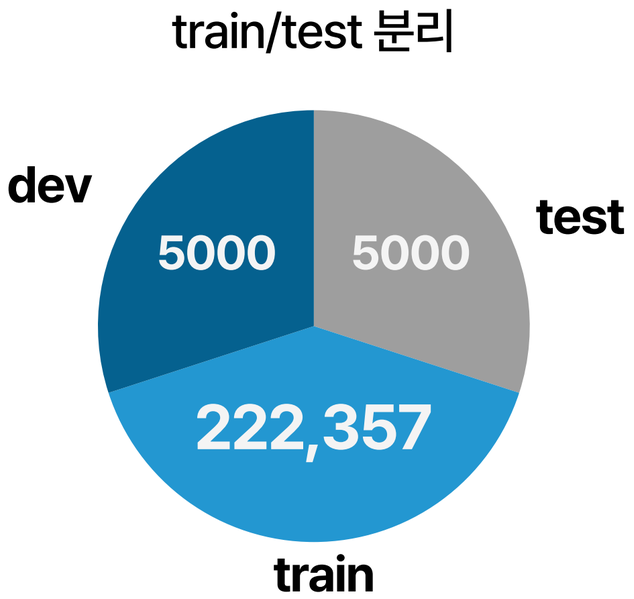
Chapter.04

모델링 KoBART fine-tuning

하이퍼 파라미터

- batch size : 16
- max_seq_len : 36
- num_ workers : 5
- ir : 5e-5
- max_epochs : 10
- warmup_ratio : 0.1
- early_stopping : 3 (val-loss 기준)

코랩 프로 A100



총 데이터 : 232,357 개

Chapter.04

모델링

val-loss를 기준으로 epoch train

A방법 최종 val-loss 0.637

B 방법 최종 val-loss 0.65808

A방법 epoch	val-loss
0	0.73281
1	0.71903
2	0.63687
3	0.66443
4	0.65491
5	0.66316

A 방법

B방법 -1 epoch	val-loss
0	1.32863
1	1.39633
2	1.24891
3	1.41850
4	1.38964

B 방법 -1

B방법 -2 epoch	val-loss
0	0.65470
1	0.69572
2	0.65808
3	0.69278
4	0.64822
5	0.70999

B 방법 -2

Chapter.04

모델링

제주도 방언 번역 예시

제주도 방언	표준어 번역
아 , 엿날 두릴 때엔 먹어낫주만은 이제 이져비언 .	아, 옛날 어릴 때엔 먹었었지만 이제 잊어버렸어.
그걸 무슨 떡이렌 골아난 ?	그걸 무슨 떡이라고 말했었어?
게난 지금 큰똘 시아방은 죽엉 낼 일포난 이제 우리 그디 갈 거 .	그러니까 지금 큰딸 시아버지는 죽어서 낼 일포니까 이제 우리 거기 갈 거.
콩찍 헹 놧다근에 그거 헤근에 그 다 밟아지면은 다 꺼냉 다라레 다 비와 놔근에 이치룩 메주 만들어근에 .	콩짚 해서 놨다가 그거 해서 그 다 밟아지면 다 꺼내서 대야에 다 비워놔서 이처럼 메주 만들어서.

SUNGKYUNKWAN

05

성능평가
performance metrics

- 평가 메트릭
- 성능 평가 결과

Chapter.05

성능평가 성능평가 metric

BLEU score

생성된 텍스트와 정답 텍스트의 overlapping 정도를 통해
평가하는 지표

ex) 정답 : 오늘 나 밥 맛있게 먹었어
생성 : 오늘 나는 밥을 맛있게 먹었습니다.
-> 의미는 유사하지만 bleu score은 낮을 것

BERTscore

생성된 텍스트의 의미적인 평가를 위한 지표
bert로 임베딩한 뒤 코사인 유사도 평균

human evaluation

최종적으로 생성된 음성의 품질을 평가하기 위해서
설문지를 구성하여 human evaluation을 진행

	모델 1(JEJURICA)	모델 2(구글 번역 음성)
발음의 정확성		
내용의 정확성		
전반적인 음질		

Chapter.05

성능평가 Bleu Score

모델명	Bleu Score
Whisper	0.20909642
KoBART	A. 0.276938 B. 0.270799
mBART	0.6366

whiper가 제주도 방언 음성에서는 성능이 살짝 떨어짐
-> 말이 빠르고 불분명하기 때문
KoBART는 **A방식**이 BLEU score가 더 높았음.

BERT score에서도 **A방식**이 성능이 더 좋았음.
-> 최종 **A방식** 모델 가중치 선택

모델명	BERT score
KoBART A	0.874817
KoBART B	0.8731798

Chapter.05

성능평가 Human Evaluation

모델명	모델 1	모델 2
발음의 정확성	142	73
내용의 정확성	*175*	*40*
전반적인 음질	115	100

총 ***43명*** 설문 참여
이 중 5명 제주도 출신

세 가지 요소에서 모두
모델 2보다 좋은 점수를 획득.
특히, **내용의 정확성**의 경우 구글 번역기보다 훨씬 더 좋은
점수를 획득.
전반적인 음질이 가장 떨어지는 편

SUNGKYUNKWAN

06

서비스 데모
Service Demo

- 서비스 구현 방법

Chapter.06

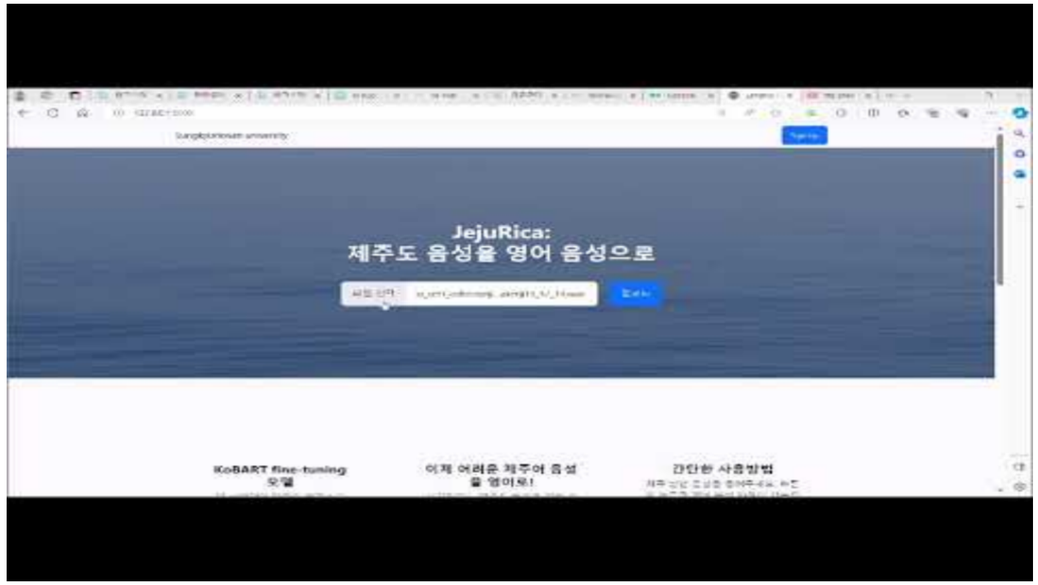
서비스 데모 Service Demo

서비스 내용

- 제주도 방언 음성 파일을 업로드하면 자동으로 영어로 번역 된 음성파일이 다운로드되도록 설정
- 음성 파일은 .wav 형식 파일만 혀용
- cpu 서버여서 전체 프로세스 약 1분 30초 소요
- flask 사용

Chapter.06

서비스데모 Human Evaluation



원본 음성 : 가이 나름 이유가 이시난 거짓말을 허겠제

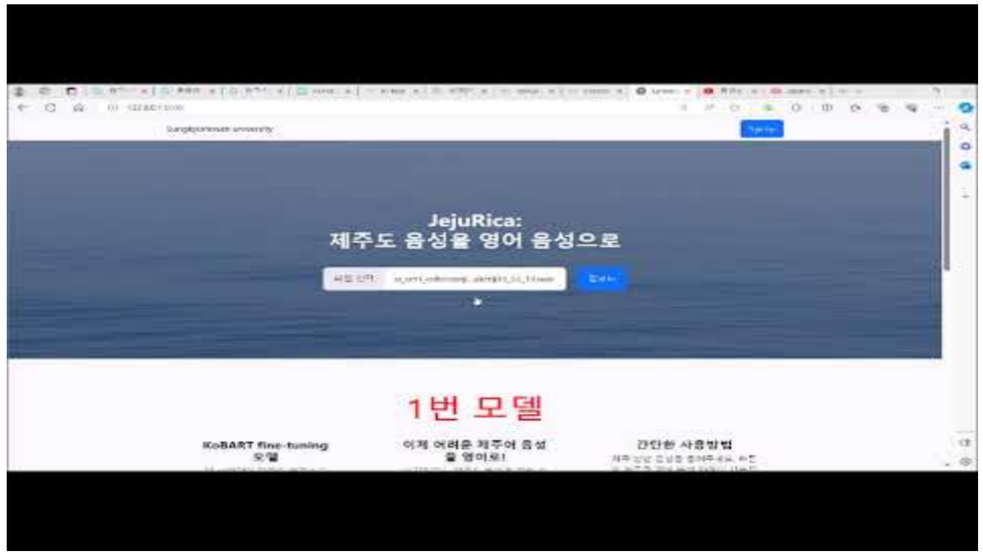
표준어 번역 : 걔가 나름 이유가 있으니까 거짓말을 하겠지

JEJURICA 영어 번역 : He'll lie for a reason.

구글 번역 음성 : Guy, I have my own reasons, but I'm willing to lie.

Chapter.06

서비스 데모 Human Evaluation



원본 음성 : 요자기 차갓집이 강보난 가시어멍이 보리 갈질 못했막 조드람서라

표준어 번역 : 요전번에 처갓집에 가서 보니까 장모가 보리 갈지를 못해서 아주 근심하고 있더라

JEJURICA 영어 번역 : Last time I went to my wife's house, my mother-in-law couldn't make barley.

구글 번역 음성 : Here, my wife’s house is surrounded by a river, and the thornfish is unable to plow the barley.

SUNGKYUNKWAN

THANK YOU
ANY QUESTIONS?