Whisper·KoBART 기반 한국어 방언-표준어 번역 시스템 설계 및 구현

이수정*, 곽대혁O
*O인하공업전문대학 컴퓨터정보공학과
e-mail: *sjlee@inhatc.ac.kr, Oeogur001@naver.com

Design and Implementation of a Whisper-KoBART-Based Korean Dialect-to-Standard Translator

Soojung Lee*, Daehyuk KwakO
*ODept. of Computer Science & Engineering, Inha Technical College

요 약

본 논문은 Whisper 기반 음성 인식, KoBART 기반 기계 번역, pyttsx3 기반 음성 합성을 통합하여 경상·제주·전라·강원 방언 발화 음성 데이터를 표준어로 변환 후 재생하는 한국어 방언 번역 시스템의 설계와 구현 과정을 제안한다. AI HUB의 방언 음성과 텍스트 데이터[3]를 정제한 뒤, 경량 Whisper-small 모델과 KoBART 번역 모델을 학습시켜 방언 음성을 텍스트로 옮기고 자연스러운 표준어로 변환하였다. 최종 시스템은 사용자가 지역을 선택하면 입력한 WAV 파일을 처리해 표준어 음성으로 들려준다. 실험 결과는 음성 인식 오차율과 번역 품질 모두 실용 수준을 만족하여, 교육 분야 및 여러 산업에 활용될 수 있을 것으로 기대된다.

▸ Keyword: 음성 인식(STT), 기계 번역(NMT), Whisper, KoBART, 딥러닝, 한국어 방언, 사투리, 표준어

I. Introduction

한국 사회의 지역 방언으로 인한 의사소통 장벽 해소를 위해 본 연구는 Whisper 기반 음성 인식, KoBART 기반 기계 번역, pyttsx3 음성 합성을 통합한 실시간 방언-표준어 변환 시스템을 설계하고 구현하였다. 이는 지역 간 소통 편의 증진뿐 아니라 방언의 보존, 콘텐츠로의 활용 등으로 확대할 수 있다.

II. Preliminaries

1. Related works

국내 방언 연구는 텍스트 번역에 집중되어 왔으며, 최근 LLM 기반의 시도가 등장했지만 방언 음성 → 표준어 음성 통합 파이프라인은 확인하기 어렵다는 점에서 선행 연구와 차별화된다.

2. Technical Background

본 연구의 파이프라인은 Whisper(STT), KoBART(NMT), pyttsx3(TTS)로 구성된다. Whisper는 경량화된 음성 인식 모델로 방언 음성을 텍스트로 변환하며, KoBART는 30,000여 문장쌍으로 미세 조정된 기계 번역 모델, pyttsx3는 표준어 텍스트를 음성으로 합성한다.

III. The Proposed Scheme

본 연구에서 제안하는 방언-표준어 변환 과정은 Fig. 1의 시스템 워크플로우로 요약된다.

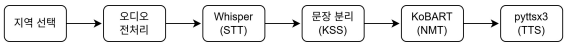

Fig. 1 시스템 워크플로우 다이어그램

1. Region Select, Model Setup

사용자가 번역을 진행할 지역을 선택하면 해당 지역의 Whisper STT 모델과 Processor, KoBART 번역 모델 및 Tokenizer, pyttsx3 TTS 엔진을 메모리에 로드하고 초기화한다.

2. Audio Chunking & Pre-Processing

입력받은 WAV 파일을 30초 단위로 분할하고 로그-멜 스펙트로그램으로 변환한다. 이 단계에서 모노 변환, 리샘플링, 정규화가 동시에 이루어진다.

3. Speech-to-Text (Whisper)

전처리된 청크를 Whisper 모델에 투입해 방언 텍스트로 변환한다.

4. Sentence Segmentation

청크별 텍스트를 하나로 합친 뒤 KSS 라이브러리를 사용해 문장 단위로 분할한다.

5. Text-to-Text Translation (KoBART)

방언 문장을 KoBART 번역 모델에 주입해 표준어 문장으로 생성한다. SentencePiece 토크나이저를 그대로 활용해 서브워드 일관성을 유지한다.

6. Text-to-Speech (pyttsx3)

생성된 표준어 문장을 pyttsx3 엔진이 즉시 합성해 출력한다. 차후 Tacotron2, FastSpeech2 등 신경망 TTS로 교체가 쉽도록 모듈 경계를 분리해 두었다.

7. Result Display & Interaction

방언-표준어 문장 쌍이 번호와 함께 표시되며, 사용자는 원하는 문장을 선택해 재생할 수 있다.

IV. Experiments

Whisper·KoBART 모델 학습을 위한 데이터 전처리, 실험 환경, 평가 지표 및 지역별 모델의 성능 분석 결과는 다음과 같다.

1. Data Source & Experimental Setup

실험은 AI HUB 지역별 방언 발화 데이터를 기반으로 수행하였다. 모델 학습 및 추론 환경은 PyTorch, Transformers, datasets로 통일했으며, KoBART 학습은 로컬 단일 GPU에서, Whisper 학습은 Google Colab T4 환경에서 진행하였다.

2. Whisper Data Pre-processing

세션 ID와 발화 구간으로 WAV를 매칭해 filepath / start / end / text 정보를 담은 manifest를 만든 뒤, 이를 16 kHz mono로 변환·슬라이스하고 Whisper Processor로 스펙트로그램 입력과 레이블을 생성해 캐시에 저장하였다.

3. Whisper Training

openai/whisper-small 모델 학습은 bf16 혼합정밀도와 그래디언트 누적 설정으로 메모리 사용을 최적화했으며, 주요 하이퍼파라미터는 learning rate 5e-6, batch 4, max steps 4000이다.

4. KoBART Data Pre-processing

JSON 파일별로 utterance의 dialect_form - standard_form을 추출해 CSV로 저장한 뒤, 특수 기호·괄호 구간을 정규식으로 제거하고 빈 문장·0.5s 미만 발화·중복·동일 문장을 필터링하였다. 그 결과 학습용 문장쌍은 경상 17만, 제주 111만, 전라 30만, 강원 92만으로 정제되었다.

5. KoBART Training

gogamza/kobart-base-v2 모델 학습은 learning rate 3e-5, batch 16, epoch 3, fp16 설정으로 파인튜닝하였다.

6. Evaluation Metrics

음성 인식 품질은 WER로, 번역 품질은 BERT Score로 평가하였다.

지역	WER(%)	BERT Score
경상	8.4	0.92
제주	7.8	0.96
전라	7.9	0.99
강원	8.1	0.99

표 1. 지역별 모델 성능

Whisper 모델은 네 지역 모두 WER 8% 내외를 기록하였으며, KoBART 모델은 검증 30 k 문장 기준 평균 BERT Score는 0.965로 방언 인식 및 번역 기능이 실용 가능함을 확인하였다.

V. Conclusions

본 시스템은 다지역 방언 음성을 실시간 표준어 음성으로 변환할 수 있고, 모델 경량화 및 데이터 서브샘플링 사용으로 자원 제약 환경에서도 효율적이다. 향후 방언 지역의 확장, 웹·모바일 GUI 개발 및 신경망 TTS 적용 등을 통해 활용 범위를 넓힐 계획이다.

References

[1] 한상민, 최은성, 이종욱, “적은 양의 병렬 말뭉치를 가진 한국어 방언 간 딥 러닝 기반 기계번역,” 한국정보과학회 학술발표논문집, 2022.

[2] 임수한 등, “CA-DiaL: 커리큘럼과 속성학습 기반의 LLM을 활용한 한국어 방언 번역,” 한국정보과학회 학술발표논문집, 2024.

[3] National Information Society Agency, “AI HUB 한국어 방언 발화 데이터셋,” 2020.