제출분야 : 빅데이터/인공지능

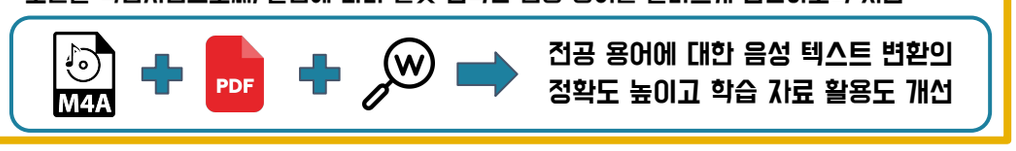
전공 용어 변환 서비스, TermCorrector

- 키워드 헌터
- 소프트웨어융합학과 / 2020105707 / 선신욱
  소프트웨어융합학과 / 2020105742 / 한지훈

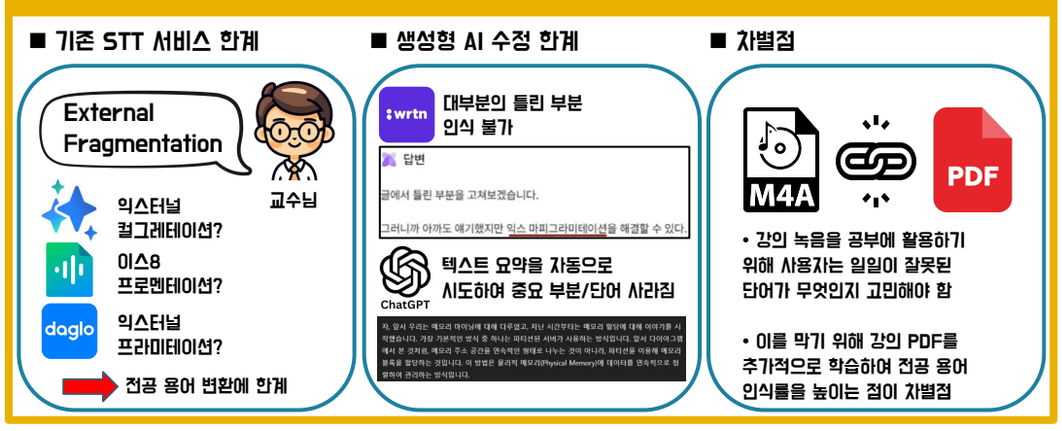
소프트웨어융합학과 / 2020105741 / 한주상

프로젝트 소개

- 기존의 Speech-to-Text(STT) 서비스에서 전공 용어의 변환 오류 발생
- 사용자에게 강의 자료(PDF)를 받아 전공 키워드를 추출하고 다양한 발음 데이터를 생성해 모델을 학습시킴으로써, 발음에 따라 잘못 인식된 전공 용어를 올바르게 변환하도록 지원



문제 정의



모델 관련

■ 키워드 추출

- 사용자가 업로드한 강의 자료(PDF)에서 영어 단어에 한해 전공 용어 키워드 추출
- 문장 단위로 토큰화 및 품사 태깅을 진행하여 NN(명사), NP(고유 명사), JJ(형용사)만 추출
- TF-IDF 방식은 일부 키워드를 놓칠 수 있는 한계가 있어 전공 용어가 아닌 키워드를 제외하는 방향으로 키워드 추출 진행

■ 데이터 증강

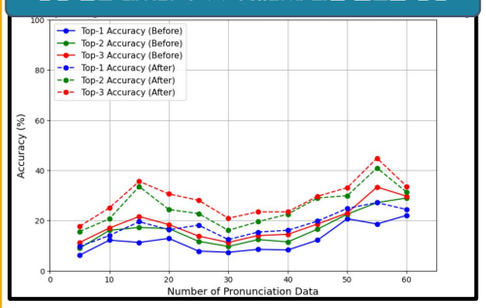
- 추출한 키워드를 g2pK 라이브러리를 사용하여 한글 발음으로 변환
- 자음과 모음에 맞는 노이즈 방식을 적용하여 글자의 초성/중성/종성 중 하나를 변형하는 방식으로 다양한 발음 변형 데이터 생성

	target	발음1	발음2	발음3	발음4	발음5
0	abstraction	애브스트랙션	캐드스트랙션	애브시트랙썰	애브스트랙션	캐브스트랙션
1	access	액쎄스	액떼스	액쎄쯔	액쎄스	익쎄스
2	address	애드레스	애브레스	애드뢰쓰	깨드레스	애드레트
3	addressable	어드레서블	어느레써블	어드레서멀	애드레서블	허드로서블
4	allocated	앨러케이티드	앨러케히티드	앨러께이티드	앨러케에티드	앨러코이테드

■ 모델 학습

- 데이터 증강 기법으로 생성한 다양한 발음 데이터를 통해 fastText 모델 학습 진행

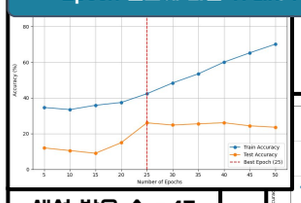
생성 발음 데이터 수 및 후처리에 따른 모델의 성능



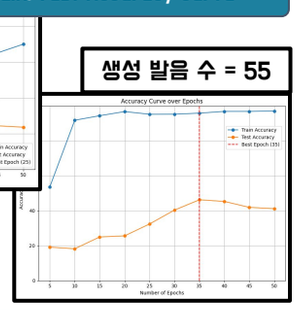
- Top-1 Accuracy, Top-2 Accuracy, Top-3 Accuracy 3가지로 모델 성능 측정
- 생성 발음 수 15개, 55개일 때 높은 정확도
- 후처리 진행 시, 정확도 향상

Epoch 변화에 따른 Train/Test Accuracy curve

생성 발음 수 = 15

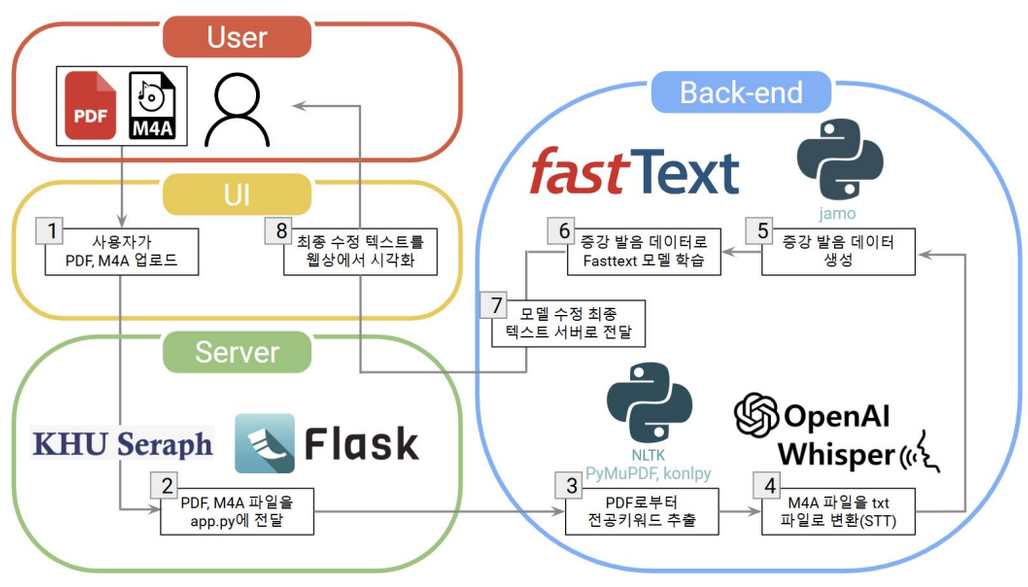


생성 발음 수 = 55



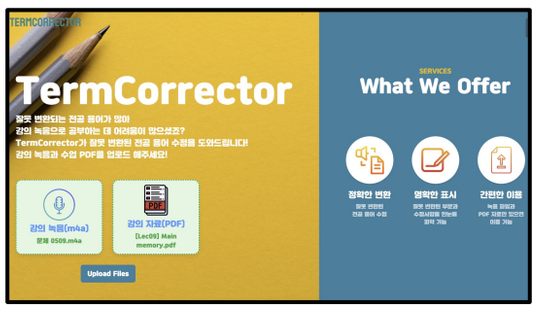
- 생성 발음 수 15의 경우 Epoch = 25에서 최적
- 생성 발음 수 55의 경우 Test Acc은 살짝 높지만 Train data에 대한 overfitting 발생 우려

프로그램 구조



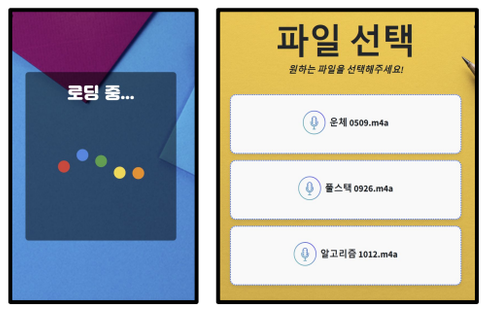
시연 과정

■ 1. 메인 페이지



- 프로젝트 기능 소개
- 사용자가 수강한 강의 녹음 파일, 강의 자료 업로드

■ 2. 로딩/파일 선택 페이지

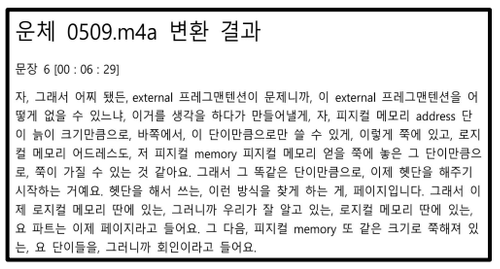


- STT 변환, fastText 모델 적용 동안 로딩
- 변환한 여러 파일 중 보고 싶은 파일 선택

■ 3. 결과 페이지

- fastText 모델이 수정한 단어 직접 확인
- 선택지 중 옳은 단어 선택 및 직접 입력 가능

■ 4. PDF 변환



운체 0509.m4a 변환 결과

문장 6 [00 : 06 : 29]

자, 그래서 어찌 됐든, external 프레그맨텐션이 문제니까, 이 external 프레그맨텐션을 어떻게 없을 수 있느냐, 이거를 생각을 하다가 만들어낼게, 자, 피지컬 메모리 address 단이 늙이 크기만큼으로, 바쪽에서, 이 단이만큼으로만 쓸 수 있게, 이렇게 쭉에 있고, 로지컬 메모리 어드레스도, 저 피지컬 memory 피지컬 메모리 얻을 쭉에 놓은 그 단이만큼으로, 쭉이 가질 수 있는 것 같아요. 그래서 그 똑같은 단이만큼으로, 이제 햇단을 해주기 시작하는 거예요. 햇단을 해서 쓰는, 이런 방식을 찾게 하는 게, 페이지입니다. 그래서 이제 로지컬 메모리 떤에 있는, 그러니까 우리가 잘 알고 있는, 로지컬 메모리 떤에 있는, 요 파트는 이제 페이지라고 들어요. 그 다음, 피지컬 memory 또 같은 크기로 쭉해져 있는, 요 단이들을, 그러니까 회인이라고 들어요.

- 최종 변환 결과를 PDF로 저장
- 웹 뿐만 아니라 로컬 파일로 확인 가능

결론

- 강의 자료(PDF)에서 전공 용어 키워드를 추출하고, 데이터 증강 기법을 활용하여 전공 용어 변환 오류 수정
- 웹상으로 수정 결과를 직관적으로 확인 가능, 바로 학습에 활용 가능
- 향후, 사용자가 유사한 domain의 여러 강의들을 업로드하여 학습하면, 이전 강의들의 전공 용어들이 모두 DB에 축적되어 전공 용어 변환 정확도 향상을 통해 더욱 유용한 학습 자료로서 활용 가능