TermCorrector: Speech-to-Text(STT)에서의 영문 용어 교정 서비스*

한지훈○, 선신욱, 한주상, 박상근
경희대학교 소프트웨어융합학과
jhhan0208@khu.ac.kr, tjstlsdnr@khu.ac.kr, han05280505@khu.ac.kr, sk.park@khu.ac.kr

TermCorrector: An English Term Correction Service for Speech-to-Text(STT)

Jihoon Han○, Shinwook Seon, Jusang Han, Sangkeun Park
Department of Software Convergence, Kyung Hee University

요 약

많은 학생들이 수업 중 강의를 녹음하여 학습에 활용하고 있으나, 교강사의 영문 용어 발음에 대한 음성 인식 정확도가 낮아 Speech-to-Test(STT) 변환 과정에서 오류가 발생하는 문제가 있다. 본 연구는 이를 해결하기 위해 강의 자료 텍스트에서 미리 영문 용어를 추출하고, 각 용어에 대해 다양한 한국어 발음 후보를 생성하여 학습하는 방법을 제안한다. 이를 통해 교강사의 발음 오류나 STT 서비스의 성능 저하로 인해 잘못 변환된 텍스트를 정확한 영문 용어로 수정할 수 있는 모델을 개발하였다. 이 모델을 활용해서 사용자가 강의 녹음 파일과 학습 자료를 업로드하면, 녹음 파일 내 잘못 변환된 한국어 텍스트를 정확한 영문 용어로 교정해 주는 웹서비스를 구현하고, 그 활용 가능성을 검증하였다.

1. 서 론

많은 대학생들이 강의를 녹음하여 학습에 활용하고 있다. 2024년 명지대학교 학생 477명을 대상으로 한 설문 조사에 따르면, 82.4%의 학생들이 강의 녹음 경험이 있다고 응답했으며, '언제든 다시 들을 수 있다'는 점과 '녹음 파일을 활용한 학습의 효율성'이 주된 이유로 확인되었다[1].

강의 녹음의 학습 효과는 기존 연구에서도 확인된다. Neil P. Morris et al.[2]은 강의 녹음이 유연한 학습, 심화 학습, 접근성과 실용성 등의 이점을 제공하여 학습에 도움을 제공함을 밝혔으며, Wendy McKenzie[3]는 강의 녹음이 학습 목표 달성과 학습해야 할 내용을 명확히 하는 데 효과적임을 강조했다.

하지만 학생들이 학습을 위해 녹음 파일을 텍스트로 변환하는 과정에서 문제가 발생하는 경우가 있다. 기존 STT 서비스는 일반적인 용어와 한글 표기로 변환된 전문 용어(예: regression → 회귀)는 비교적 정확히 인식하는 반면, 영문 기반의 전문 용어를 한국어로 변환할 때는 발화자의 외래어 발음 차이로 인해 정확한 용어 인식에 어려움이 있다[4]. 예를 들어, "external fragmentation"이라는 용어가 "이스8 프로멘테이션"와 같이 변환될 수 있다. 발화자의 발음이나 STT 서비스의 한계로 인해, 부정확한 한국어 텍스트 변환이 발생할 수 있으며, 이는 학생들이 강의 녹음 파일을 학습에 활용하는 데 큰 방해 요소로 작용한다.

이 문제를 해결하기 위해 전문 용어 인식 정확도를 개선하기 위한 다양한 연구가 진행되었다[5, 6, 7, 8]. 하지만 기존의 연구는 전문 용어의 인식 정확도를 개선했지만, 대규모의 데이터 셋을 수집해야 하거나 데이터 전처리 및 모델 학습 과정에서 오랜 시간이 소요된다는 한계가 존재한다.

본 연구에서는 기존 STT 서비스에서 영문 텍스트의 변환 오류가 빈번하다는 점을 고려하여, 사용자가 업로드한 강의 자료에서 영문 용어를 추출하고, 이를 기반으로 강의 녹음 음성을 텍스트로 변환 시 잘못 변환되는 용어를 찾아 교정하는 인공지능 모델을 개발했다. 해당 모델의 활용성을 확인하기 위해, 사용자가 강의 녹음 파일(m4a)과 강의 자료 파일(PDF)을 업로드하면, 틀린 텍스트로 변환된 용어를 교정해 주는 웹 서비스를 개발해서 활용 가능성을 확인했다.

2. 관련 연구

전문 용어 인식 성능을 개선하기 위한 연구가 활발히 이루어지고 있다. Suh et al.[5]은 음성 인식 모델인 Whisper[1]의 Decoder에 LLM으로 생성한 간략한 설명을 추가하고 학습하여 전문 용어의 인식 정확도를 개선하는 방법을 제안했다. Anh & Sy[6]는 LLM과 프롬프트를 모두 활용하여 맥락 정보를 효율적으로 활용하는 새로운 음성 인식 모델 아키텍처를 제안했다. 텍스트 Encoder를 고정하고 어댑터를 활용하여 문맥 정보를 반영하였으며, LLM 기반 재예측을 통해 인식 정확도를 높였다. Mani et al.[7]는 기계 번역 모델을 이용하여 음성 인식 시스템이 생성한 텍스트의 오류를 수정하는 방법을 제안했다. 구체적으로 Sequence-to-Sequence 및 Transformer 모델을 사용하여 오류가 포함된 음성 인식 출력 결과와 정답 텍스트 쌍을 학습시킴으로써 전문 용어의 변환 오류를 수정하였다. 음성 인식 시스템 자체를 개선하기보다 데이터 셋의 품질과 양을 늘려 성능 개선을 이룬 연구도 진행되었다. Jia[8]는 대규모의 인공 데이터와 실제 데이터를 수집하여 고품질 주석 데이터를 생성하고, 이를 음향 모델과 언어 모델을 학습시켜 전문 용어의 인식 정확도를 개선하는 방법을 제안하였다.

* "본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025년도 SW중심대학사업의 결과로 수행되었음"(2023-0-00042)

1) https://github.com/openai/whisper

기존 연구들은 주로 Fine-tuning을 통해 전문 용어의 음성 인식 성능을 개선하였다. Fine-tuning이 특정 언어의 인식률을 향상시킬 수 있었지만, 다른 언어의 인식률을 저하시킬 수 있다는 한계가 있어[9], 본 연구에서는 Fine-tuning 없이 영문 용어의 변환 오류 문제를 해결하는 방법을 제안한다. 또한, 실시간으로 업로드 되는 강의 녹음의 도메인이 예측 불가능해 모든 분야의 대규모 데이터 셋 확보가 어려운 상황을 고려하여 사용자가 업로드 한 강의 자료(PDF)에서 키워드를 추출하고, 데이터 증강 기법을 활용하여 정확한 텍스트 변환을 위한 자체 데이터 셋을 구축했다. 이를 기반으로, 강의 녹음 파일(m4a)을 활용한 사용자의 학습에 도움이 될 수 있도록 한국어 단어를 정확한 영문 용어로 변환할 수 있는 웹서비스를 개발해서 그 활용성을 검증했다.

3. 전문 용어 기반 STT 교정 모델 개발

3.1 교정 대상 용어 선정

강의 자료 파일 가공: 강의 자료 파일(PDF)이 주어지면 영문 용어 키워드를 추출하고, 각 영문 용어마다 표준 한국어 발음 텍스트를 매칭한다. 이를 위해, 파일에서 텍스트를 추출한 후 문장을 단어 단위로 토큰화(tokenization)하고, 의미 있는 단어를 선별하기 위해 NLTK 라이브러리를 활용하여 명사(NN), 고유명사(NNP), 형용사(JJ) 품사에 해당하는 단어만 추출한다. 추출된 단어는 모두 소문자로 변환하여 중복 처리를 최소화하고, 표제어 추출(lemmatization)을 적용하여 같은 의미를 가진 단어들을 하나로 통합한다. 그리고 NLTK[2]에서 제공하는 영어 불용어(stopword) 리스트와 한 글자로 구성된 단어를 제거하고, 정규 표현식으로 이메일 및 링크 등과 같이 학습에 필요한 용어와 관련 없는 단어를 식별하여 모두 제거한다. g2pK 라이브러리를 사용하여, 남은 텍스트마다 표준 한국어 발음 텍스트를 생성한다 (예: external → 익스터널).

강의 녹음 파일 가공: 교수자의 강의 녹음 파일(m4a)이 주어지면 whisper를 활용해 해당 녹음 파일을 한국어 텍스트로 변환하고, KoNLPy[3] 라이브러리의 Okt 형태소 분석기를 사용하여 명사, 고유명사, 형용사만 추출했다. 그리고 유사도 계산 라이브러리인 difflib[4]를 활용해서 강의 자료 파일(PDF)에서 생성한 영문 용어 키워드의 한국어 발음 텍스트와 비교해서 유사한 텍스트를 추출한다 (유사도 기준 0.7 이상).

예를 들어, 강의 자료 파일(PDF)에 'fragmentation'이라는 용어가 있다면 이는 g2pK 라이브러리를 통해 표준 한국어 발음인 '프라그멘테이션'으로 변환된다. 강의 녹음 파일에서 추출된 한국어 텍스트 중 '쁘라그멘테이션'이라는 용어가 있다면 '프라그멘테이션'과 '쁘라그멘테이션'의 유사도가 0.769로 계산되므로, 강의 녹음 파일에서 추출된 '쁘라그멘테이션' 용어는 교정 대상 용어로 선정된다.

3.2 데이터 증강 및 모델 학습

강의 자료 파일(PDF)에서 추출한 영문 텍스트를 표준 한국어 발음 텍스트로 변환한 다음, 교수자의 녹음 파일에서 추출한 텍스트와 비교할 때 실제 영문 용어와 교정 대상 용어가 잘못 매칭되는 경우가 있음을 확인했다. 예를 들어, 강의 자료 파일에서 추출한 'Fragmentation'을 표준 한국어 발음으로 변환하면 '프라그멘테이션'이 된다. 이 텍스트를 교수자의 강의 녹음 파일에서 추출한 텍스트와 비교할 때, '쁘라그멘테이션'과 매칭이 되어야 하는데, '세그멘테이션(Segmentation)'과 잘못 연결되는 오류가 발생할 수 있다.

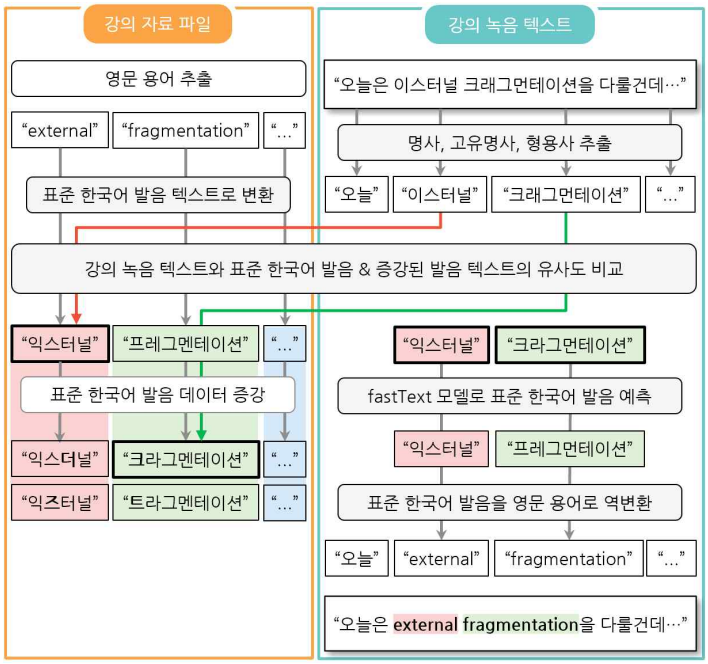

그림 1. fastText 모델을 활용한 영문 용어 교정 예시

이러한 문제를 해결하기 위해, 한국인 영어 학습자의 발음 오류 패턴[11, 12]을 반영하여 데이터 증강 기법을 적용하였다. 표준 한국어 발음 텍스트를 기반으로 유사한 자음·모음을 활용해 다양한 변형 텍스트를 생성함으로써, 발음 변형에 따른 인식 오류를 줄이고자 하였다. 강의 자료 파일(PDF)의 용어 'fragmentation'이 g2pK 라이브러리를 통해 표준 한국어 발음 '프라그멘테이션'으로 변환되었을 때, 이 용어에서 랜덤하게 'ㅡ'라는 글자를 선정한다. 그리고 'ㅍ'과 유사한 'ㅋ', 'ㅌ'을 활용해 '크라그멘테이션', '트라그멘테이션'이라는 새로운 텍스트가 생성된다. 이 새로운 텍스트들과 교수자의 발음 텍스트를 비교하면 이와 가장 유사한 교수자의 발음 텍스트(예: '크래그먼테이션')를 찾아낼 수 있다. 만약 교수자가 발음한 'Fragmentation' 용어가 '크래그먼테이션'으로 변환된 경우, 표준 한국어 발음인 '프라그멘테이션'과는 유사도가 상대적으로 높지 않지만, 증강해서 새롭게 생성한 텍스트인 '크라그멘테이션'과는 상대적으로 유사도가 높다. 그러므로 '크래그먼테이션'이라는 용어가 교정 대상 용어로 선정되고 프라그멘테이션(Fragmentation)으로 교정할 수 있다. 만약 데이터를 증강하지 않았다면 '세그멘테이션(Segmentation)'과 가장 유사하다고 판단되어 잘못된 교정이 발생할 수 있다.

교정 대상으로 선정된 용어를 정확한 영문 용어로 변환하기 위해 페이스북의 워드 임베딩 및 텍스트 분류 라이브러리 fastText[10]를 사용했다. 증강된 데이터로 학습된 fastText 모델은 강의 녹음 파일에서 변환된 텍스트 중 교정 대상으로 선정된 용어들(예: '이스터널', '크래그먼테이션')을 입력으로 받아, 각각 'external', 'fragmentation'과 같은 정확한 영문 용어로 교정할 수 있다 [그림 1].

2) https://www.nltk.org/
3) https://konlpy.org
4) https://docs.python.org/ko/3.13/library/difflib.html

4. 웹서비스 개발

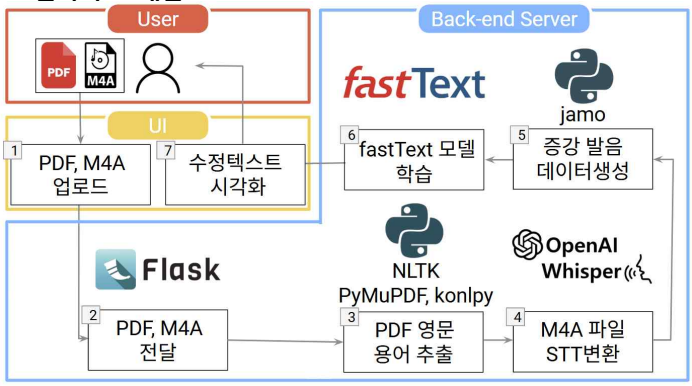

그림 2. TermCorrector 아키텍처

본 연구에서는 위에서 개발한 STT 교정 모델을 활용해서, 사용자들이 교수자의 녹음 파일을 텍스트로 변환해서 공부할 때, 잘못된 한국어 발음이 아닌 교정된 정확한 영문 용어로 학습할 수 있는 웹서비스 TermCorrector[5]를 개발했다. TermCorrector의 아키텍처는 [그림 2]와 같다.

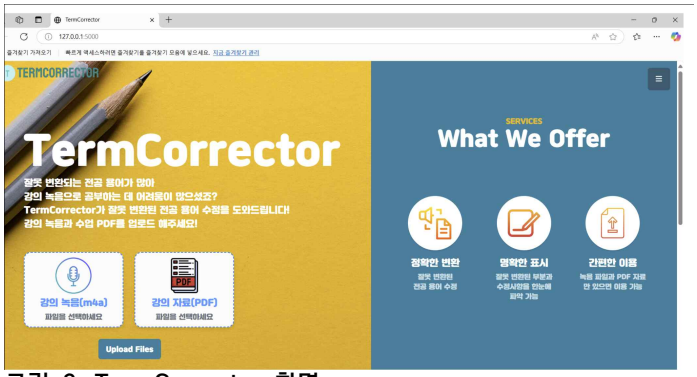

그림 3. TermCorrector 화면

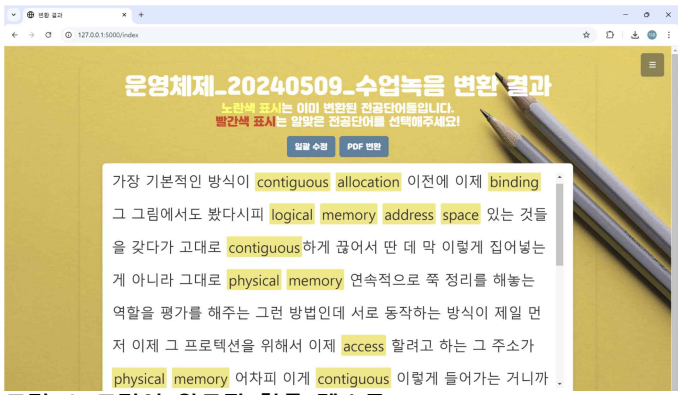

그림 4. 교정이 완료된 최종 텍스트

사용자는 웹서비스 메인 화면에서 강의 자료 파일(PDF)과 강의 녹음 파일(m4a)을 업로드할 수 있으며 [그림 3], 강의 녹음 파일을 업로드하면 서버에서 이 파일을 텍스트로 변환한다. 그리고 강의 자료 파일을 기반으로 교수자의 강의 녹음 파일에서 잘못 변환된 한국어 텍스트를 정확한 영문 용어로 교정한다. 사용자는 웹서비스에서 교정이 완료된 최종 텍스트를 바로 확인할 수 있다 [그림 4].

5) https://youtu.be/J-eTbh80qko (Demo Video)

5. 결론

본 연구에서는 사용자가 업로드한 강의 녹음 파일과 강의 자료 파일을 기반으로, STT 변환 과정에서 잘못 인식된 한국어 텍스트를 정확한 영문 용어로 교정해 주는 모델을 개발하고, 이를 바탕으로 웹서비스를 구현하였다. 사용자의 강의 자료에서 영문 용어를 추출하고, 각 용어에 대한 다양한 한국어 발음 후보를 생성하여 학습에 활용함으로써, Fine-tuning 없이도 변환 오류를 효과적으로 보완했다는 점에서 기존 연구와 차이가 있다. 또한 사용자는 교정 결과를 직접 선택하거나 수정할 수 있으며, 최종 결과를 PDF로 저장해 학습 자료로 활용할 수 있다. 향후에는 강의 자료에서 추출한 키워드를 DB에 축적하여 활용함으로써, 다양한 분야에서의 활용성과 변환 정확도를 높이고, 학습 지원 도구로서의 활용성을 더욱 확장할 수 있을 것으로 기대된다.

6. 참고문헌

[1] 황성용. "강의 녹음, 효율적인 학습을 위해 필요한 논의 〈1133호〉", 명대신문, 20024.09.30

[2] Neil P. Morris, et al. "Lecture recordings to support learning: A contested space between students and teachers", Computers & Education, 140, 103604, 2019

[3] Wendy McKenzie. "Where are audio recordings of lectures in the new educational technology landscape", Proceedings ascilite Melbourne, 2008

[4] 강민순. "영어 외래어 한글 표기가 영어 발음에 미치는 영향", 국제차세대융합기술학회, 2022, 1708-1715, 2022

[5] Jiwon Suh, et al. "Improving Domain-Specific ASR with LLM-Generated Contextual Descriptions", Interspeech, 2024

[6] Nguyen Manh Tien Anh, Thach Ho Sy. Improving Speech Recognition with Prompt-based Contextualized ASR and LLM-based Re-predictor", Interspeech, 737-741, 2024

[7] Anirudh Mani, et al. "Error Correction and Domain Adaptation Using Machine Translation", ICASSP, 6344-6348, 2020

[8] Yanan Jia. "A Deep Learning System for Domain-specific Speech Recognition", arXiv:2303.10510, 2023

[9] 오창한, et al. "대형 사전훈련 모델의 파인튜닝을 통한 강건한 한국어 음성인식 모델 구축", 한국음성학회, 75-82, 2023

[10] Piotr Bojanowski, et al. “Enriching Word Vectors with Subword Information”, TACL, 2017.

[11] 유혜배, 윤한나. "한국인 성인 영어학습자의 대화상에 나타난 분절음 발음오류 연구", 인문학연구, 185-212, 2009

[12] 박시균. "한국인 영어 학습자의 발음 오류 원인 분석과 교육방법 모색", 한국언어학회, 113-143, 2004