2023년 정보컴퓨터공학부 전기 졸업과제 중간 보고서

지식베이스를 활용한 외래어 중심의 G2P 연구

분과: A(인공지능)

지도교수: 권혁철

팀 이름: RNG아니조

팀원: 201724412 권민규

201524410 고상현

201824468 박건우

1. 요구 조건 및 제약 사항 분석

1.1 과제 배경 및 요구 조건

- 현재 상용 TTS들의 발음을 조사해 본 결과, 외래어 부분에서 한국어로 발음을 변환하는데 있어서 미흡한 부분이 존재하는데 이를 개선하고자한다.

ex1) 네이버 '파파고'의 경우

→ "WCDMA가 도입돼 통화 중 상대방의 모습을 생생히 볼 수 있고 원격 화상회의도 일반화될 전망이다." 원래 발음: 더블유시디엠에이 / tts 변환 발음 : 크드마

ex2) 마이크로소프트 'bing'의 경우

→ "궉 박사는 현재 유체역학 학술지인 유체물리학지의 편집장과 미국 UCLA석좌교수로 활동하고 있다." 원래 발음: 유시엘에이 / tts 변환 발음 : 유시알에이

- 성능을 개선하기 위해서는 기분석 사전을 기반으로 하는 지식베이스 방식이 요구된다.

- 기분석 사전에 존재하지 않는 단어를 변환하기 위해서 딥러닝 모델 활용도 요구된다.

1.2 과제 제약 사항 분석

- 모든 외래어에 대하여 db를 구축하고 발음을 처리하기에는 고려해야 할 단어의 개수가 너무 많아서 현실적으로 어려움이 있다.

→ 사전에 '외래어'로 등재된 단어와 영어 알파벳으로 이루어진 단어에 대해서만 범위를 한정하기로한다.

- 사전에 등록되지 않은 외래어가 다수 존재한다.

  → 우리말샘 사전과 IPA 사전을 기반으로 구축한 db에 해당 단어가 없을 경우, IPA 발음 규칙을 활용하여 한국어로 변환하거나, 딥러닝 모델을 통한 변환을 진행하기로한다.

2. 설계 상세화 및 변경 내역

2.1 기분석 사전 구축

입력으로 들어오는 외래어를 올바른 한국어 발음으로 변환하기 위해서 사용될 기분석 사전을 구축한다. 우선 우리말샘 사전에서 전체 단어 데이터들을 다운 받아 필요한 형태로 전처리한다. 전처리된 데이터를 어휘에 따라서 구분하여 DB로 각각 저장한다. ----(그림1)

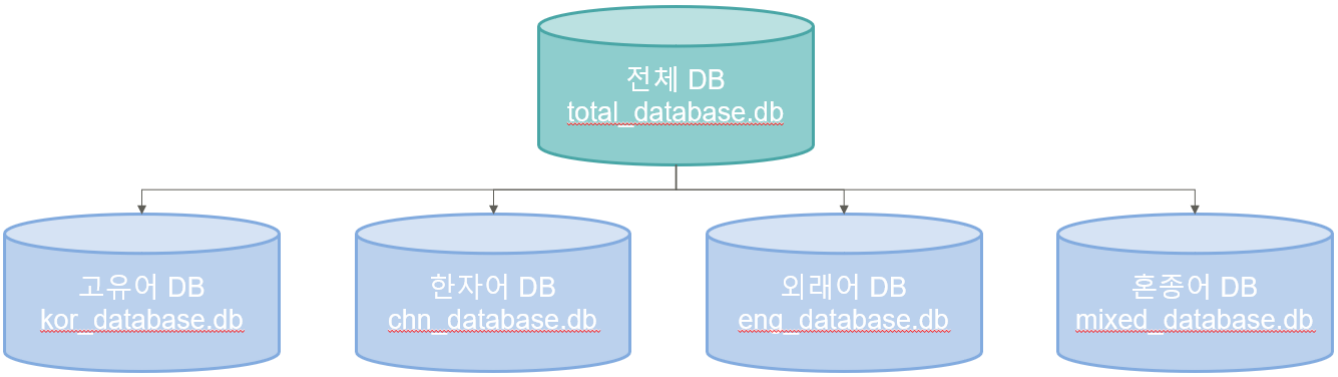


그림1 전처리된 DB 구조

전처리된 전체 DB에 존재하는 모든 conju_list(단어의 활용형) 의 표기와 발음을 추출하여 새로운 활용어 DB인 conju_db 를 생성한다. ----(그림2)

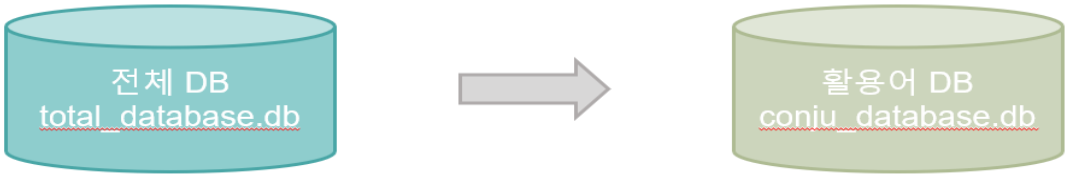


그림2 활용어 DB

ID	word	word_type	word_unit	conju_list	pronun_list	sense_no	sense_type	pos	origin_lang	origin_lang_type
57253	챗봇	외래어	어휘	None	챗봇	001	일반어	명사	chatbot	영어
57254	챠밍 포인트	외래어	구	None	챠밍포인트	001	일반어	명사	▼charming point	영어
57255	챠트	외래어	어휘	None	챠트	001	일반어	명사	chart	영어
57256	처널리즘	외래어	어휘	None	처널리즘	001	일반어	명사	churnalism	영어

그림3 외래어 DB 속 저장된 데이터 형식 예시

우리말샘 사전에 존재하는 단어의 개수는 총 1,164,952개이다. 위의 전처리 과정을 거지고 난 뒤, 남게 되는 단어의 개수는 779,429개가 된다. 본 과제에서는 81,539개의 외래어 DB를 주로 사용한다.

index	고유어	외래어	한자어	혼종어
어휘	254,229	51,413	322,188	125,807
구	7,199	42,735	219,929	125,324
명사	136,849	50,862	50,862	63,398
대명사	559	0	258	16
수사	105	0	21	0
조사	590	0	0	0
동사	58,047	0	0	46,688
형용사	19,337	0	0	9,996
관형사	397	0	89	4
부사	28,425	4	709	3,348
감탄사	1,493	22	29	17
접사	462	0	432	2
의존 명사	760	503	388	99
보조 동사	108	0	0	0

그림4 우리말샘 사전에 존재하는 전체 단어에 대한 통계

	개수
전체	779,429
고유어	80,767
외래어	81,539
한자어	411,718
혼종어	205,405

그림5 전처리 과정 후 필터링 된 단어의 개수

2.2 변경사항

사전에 DB를 구축할 때, sense number 가 1인 경우 (사전에서 단어를 검색했을 때, 가장 첫 번째로 나오는 뜻)만 가지고 데이터를 처리 하였기 대문에, sense number 가 1이 아닐 경우, DB 에 존재하지 않기 때문에 발음열 매칭이 불가능하다.

Ex)

엔(en[円])

「001」일본의 화폐 단위. 기호는 ￥.

엔(N / n)

「003」영어 알파벳의 열네 번째 자모 이름.

위처럼, 엔(N/n) 은 sense number 가 3이기 때문에 DB에 저장되지 않는다. 따라서

"자석은 N극과 N극이 만나면 서로 밀어냅니다." 과 같은 문장의 N을 제대로 변환할 수 없다.

이러한 문제를 해결 하기 위해, sense number 가 1인 단어만 DB에 저장하는 것이 아닌 그림3 에서 볼 수 있는 origin_lang 을 기준으로 단어를 DB에 저장하기로 한다.

또한, 우리말샘 사전을 기반으로 구축한 DB에 존재하지 않으면서, 대문자로만 이루어진 단어의 경우, 알파벳 그대로 끊어서 발음할 수 있지만, DB에 존재하지 않고, 대문자로만 이루어지지 않은 경우에는 문제가 발생한다.

Ex) "아이폰은 iOS 운영체제를 탑재하여 저희 Hello World사의 App과 호환되지 않습니다."

iOS[아이오에스], Hello[헬로], World[월드]는 DB에 존재하지 않고 대문자로만 이루어지지 않은 외래어

위와 같은 문제를 해결하기 위해서 딥러닝 모델 구축하기로 한다.

3. 갱신된 과제 추진 계획

<table>
<tr><th colspan="5">6월</th><th colspan="4">7월</th><th colspan="5">8월</th><th colspan="4">9월</th></tr>
<tr><td>1주</td><td>2주</td><td>3주</td><td>4주</td><td>5주</td><td>1주</td><td>2주</td><td>3주</td><td>4주</td><td>1주</td><td>2주</td><td>3주</td><td>4주</td><td>5주</td><td>1주</td><td>2주</td><td>3주</td><td>4주</td></tr>
<tr><td colspan="9">기분석 사전 구축</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="4">사전 검증 및 수정</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="2">중간 보고</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="4">추가 DB 구축 및
학습 데이터 수집</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="4">DB 검증 및 딥러닝
모델 생성</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="4">모델 검증 및 최적화, 디버깅</td><td></td></tr>
</table>

4. 구성원별 진척도

이름	진척도
권민규	기분석 사전 구축 및 추가 데이터 수집/전처리
고상현	학습 데이터 수집 및 발음 변환 규칙 구현
박건우	구축된 사전 DB 검증 및 과제에 적합한 모델 조사

5. 보고 시점까지의 과제 수행 내용 및 중간 결과

외래어가 포함된 학습 데이터를 앞에서 구축한 DB를 이용하여 얼마나 매칭되는지 실험해 보았다. 학습에 사용된 총 문장의 개수는 1865개이다.

학습 데이터 총 문장수	1865개
문장 속 외래어 개수	2260개
중복을 제외한 외래어 개수	469개

그림6 학습 데이터 속 외래어 통계

Database 내에 존재	Database 내에 존재하지 않음	
	대문자로 이루어짐	대소문자 혼합 or 소문자
248개	187개	34개

그림7 학습 데이터를 DB와 비교한 결과

※ DB와 학습 데이터 매칭 방식

1. 문장에서 영어 단어를 뽑아와서 DB에서 확인 (대소문자 구분)

2. 대소문자 구분해서 없으면 대소문자 구분 없이 재확인

3. 1, 2번 결과 DB에 존재하지 않으면 알파벳 단위로 끊어서 읽기

기분석 사전에 최대한 많은 외래어를 저장하려고 했으나 앞선 통계에서 볼 수 있듯이, DB에 존재하지 않는 단어들이 아직 많이 남아 있다. 이를 보완하기 위해, 활용할 딥러닝 모델을 지속적으로 조사하는 중에 있으며, 현재까지 조사한 결과로는, Transformer 모델이나 LSTM 모델을 사용하고자 한다.

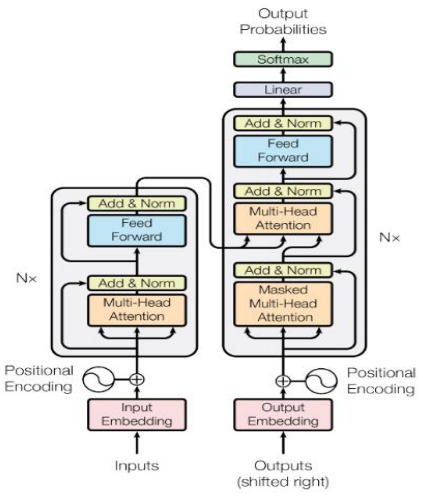
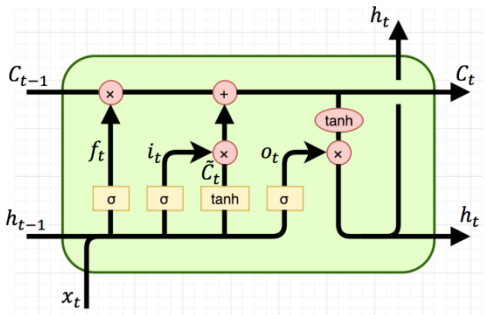


Figure 1: The Transformer - model architecture.



(a) Long Short-Term Memory