Pusan National University
Computer Science and Engineering
Technical Report 2023-10

지식베이스와 딥러닝을 통한 영어 – 한글 Grapheme-to-Phoneme 모델 성능 향상 연구

201724412 권민규

201524410 고상현

201824468 박건우

지도교수 권혁철 교수

목 차

1. 서론

1.1. 연구 배경

텍스트를 음성으로 변환하는 TTS(Text-to-Speech) 기술은 오디오북, 네비게이션, 시각 장애인 보조 도구 등 다양한 분야에서 활용되고 있고, 오랜 시간에 걸쳐 발전되어왔다. 그러나 한국어 발음 규칙과, 영어 발음 규칙에는 차이가 존재하기 때문에, 현재 상용화된 대중적인 TTS 중에 영단어를 제대로 된 한국어 발음으로 변환하지 못하는 경우가 여전히 다수 존재한다. 따라서, 이 부분을 개선하면 언어 교육 애플리케이션에서 학습자들의 정확한 발음 연습을 돕거나, 시각 장애인과 같이 읽는데 어려움을 겪는 사람들에게 도움을 제공하는 등 긍정적인 기대 효과가 예상되기에 본 과제의 주제를 이와 같이 선정하였다.

1.2. 기존 문제점

본 과제를 수행하기에 앞서 상용 TTS 들의 영어 → 한국어 발음 변환 성능을 조사해 보기로 하였다. 국내에서 가장 이용자가 많을 것으로 예상되는 네이버의 '파파고'와 마이크로소프트의 'Bing'을 위주로 테스트를 진행하였다.

[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**

"트랜스포머 모델 관련 논문을 읽고 multiheadattention 에 대해 정리한 내용입니다."

→ 원래 발음: 원래 발음: **멀티헤드어텐션** / TTS 변환 발음 : **멀티헤이버트텐션**

[ex2] 마이크로소프트 'Bing'의 경우

"권 박사는 현재 유체역학 학술지인 유체물리학지의 편집장과 미국 UCLA석좌교수로 활동하고 있다."

→ 원래 발음: **유시엘에이** / TTS 변환 발음 : **유시알에이**

위의 예시에서 볼 수 있듯이, 일부 영단어의 경우, 부정확한 발음으로 변환하는 것을 확인할 수 있었다. 이는 대부분의 TTS가 발음을 변환하는 과정에서 딥러닝 기술을 중점적으로 사용하기 때문인 것으로 추측된다.

1.3. 연구 목표

우리는 이 부분에서 차별점을 두기 위해, 지식베이스와 딥러닝을 결합한 방식의 시스템을 구축하여 본 과제를 수행하고자 한다. 공개 되어있는 영단어 발음 데이터를 수집하여, 기분석 사전을 구축한 뒤, 입력된 단어가 기분석 사전에 존재하는 단어일 경우, 사전에 등록된 발음으로 바로 변환하고 그렇지 못한 경우에 딥러닝 기술을 활용하여 변환하는 것을 목표로 한다. 착수 단계에 구상하였던 최종 시스템의 흐름도는 아래의 그림과 같다.

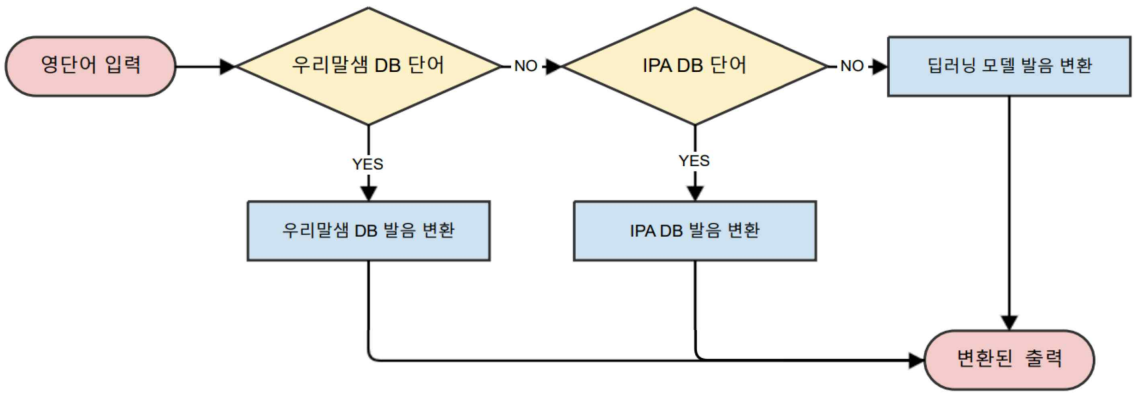


그림 초기에 설계한 시스템에 대한 Flowchart

딥러닝 모델은 LSTM을 기반으로 한 seq2seq 모델을 먼저 활용해보기로 결정하였다. 인코더 부분은 양방향 LSTM을 사용하여 주변 정보를 균형 있게 담도록 하고, teacher forcing 방식을 도입하여 빠르고 정확한 예측을 할 수 있는 방식으로 디코딩을 진행하면 정확도 높은 모델을 만들 수 있을 것이라고 예상하였다.

2. 연구 배경

2.1. 데이터 수집

본 과제에서 최우선적으로 활용할 데이터는 우리말샘 사전으로 선정하였다. 기존의 사전들은 여러 가지 제약 사항이 있어 단어를 제한적으로 수록하였으나, 우리말샘 사전은 신어, 전문 용어를 포함한 다양한 어휘를 수록하고 있다. 또한 각 분야의 전문가들이 참여하여 편집된 내용을 감수하도록 하여 정보의 신뢰도도 보장되는 점이 데이터로 선정하게 된 이유이다. 우리말샘 사전에는 총 1,164,952개의 단어가 존재하는데, 이 중에 81,593개의 외래어만 추출하여 본 과제에 필요한 형태로 전처리 하였다.

그

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

에

존재하는 전체 단어에 대한 통계

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

그림 3 필터링 된 단어의 개수

아래의 그림은 우리말샘 DB에 저장된 데이터의 예시를 보여주는 표이다.

ID	word	word_type	pronun_list	sense_no	origin_lang	origin_lang_type
39214	유로	외래어	유로	008	Euro	영어
50102	오거	외래어	오거	003	augur	영어
59347	하트	외래어	하트	004	chart	영어

그림 4 우리말샘 DB 속 저장된 데이터 형식 예시

2.2. 데이터 개선 사항

초기에 DB를 구축할 때, sense number 가 1인 경우 (사전에서 단어를 검색했을 때, 가장 첫번째로 나오는 뜻)만 DB에 저장하도록 하였기 때문에, sense number 가 1이 아니라는 이유로 DB에 등록되지 못하는 단어가 많아 정확한 발음열 매칭이 불가능하였다.

[ex] 엔(en[円])

「001」일본의 화폐 단위. 기호는 ￥.

[ex] 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 없었다.

이러한 문제점을 개선 하기 위해, sense number 가 1인 단어만 DB에 저장하는 것이 아니라 그림4 에서 볼 수 있는 'origin_lang' 을 기준으로 단어를 DB에 다시 저장하였다.

3. 연구 내용

3.1. 데이터 추가 수집

우리말샘 사전으로 생성한 DB만을 이용하여 영어 단어의 발음을 변환하기에는 부족한 부분이 아직 많이 존재하였다. 그래서 시스템의 정확도를 향상시키기 위해서 추가적인 DB를 구성하는 것이 필요하다고 느끼게 되었고, 자료를 조사하던 중, IPA 사전을 찾게 되었다. IPA란 국제 음성 기호로, 모든 언어의 음성을 표기하기 위한 표준화된 기호 체계이다. 이 기호 체계를 활용하여, 영어 단어들의 발음을 IPA 기호로 표현한 사전이 있는데 이것이 IPA 사전이다. IPA 사전은 우리말샘 사전 보다 훨씬 많은 영단어를 포함하고 있기 때문에 이것을 추가적인 DB로 활용하기로 결정하였다.

3.1.1. IPA 변환 규칙

IPA 사전에는 영단어의 대한 발음이 IPA(국제음성기호)로 표시되어 있어, 이를 한국어 발음으로 변환하는 과정이 필요하였다. 그래서 국립 국어원의 외래어 표기법 조항을 참고하여 IPA 기호를 한글 발음으로 변환하기로 하였다. 국립 국어원에서 공표한 '외래어 표기법 제 3장 영어 표기 세칙' 을 적용하여 변환을 시도하였으나, 국립 국어원에서 고시한 규칙은 영어 알파벳을 기준으로 작성 되어 있기 때문에, IPA 기호와 정확히 매칭하는데 어려움이 있었다. 그래서 우리는 해당 알파벳과 최대한 비슷한 IPA 기호와 매칭하여 발음을 변환하기로 결정하였다.

1. 외래어 표기법 제 3장 표기세칙 제 1절 영어의 표기 제 9항 1호

[w]는 [wə], [wɔ], [wou]는 '워', [wɑ]는 '와', [wæ]는 '왜', [we]는 '웨', [wi]는 '위', [wu]는 '우'로 적음.

'wə' : '워', 'wɔ' : '워', 'wou' : '워', 'wɑ' : '와', 'wæ' : '왜', 'we' : '웨', 'wi' : '위', 'wu' : '우'

Word[wə:d] 워드　want[wɔnt] 원트　woe[wou] 워　wander[wɑndə] 완더

2. 외래어 표기법 제 3장 표기세칙 제 1절 영어의 표기 제 9항 2호

자음 뒤에 [w]가 올 때에는 두 음절로 갈라 적되, [gw], [hw], [kw]는 한 음절로 붙여 적음.

'ɡw':'구', 'hw': '후', 'kw': '쿠',

Swing[swiŋ] 스윙　twist[twist] 트위스트

3. 외래어 표기법 제 3장 표기세칙 제 1절 영어의 표기 제 9항 3호

반모음 [j]는 뒤따르는 모음과 합쳐 '야', '얘', '여', '예', '요', '유', '이 ' 로 적음.

다만, [d], [l], [n] 다음에 [jə]가 올 때에는 각각 '디어', '리어', '니어'로 적음.

'djə':'디어', 'ljə': '리어', 'njə': '니어',

yard[jɑ:d] 야드　yank [jæŋk] 얭크　yearn[jə:n] 연　yellow[jelou] 옐로

3.1.2. IPA DB 와 우리말샘 DB 비교

앞서 언급한 발음 규칙의 정확도를 알아 보기 위해서, IPA DB와 우리말샘 DB에 동시에

존재하는 단어의 발음을 비교하여 정확하게 일치하는지 확인해보았다. IPA DB와 우리말샘 DB에 공통으로 존재하는 단어는 총 12,443개 였는데 그 중 발음이 정확히 일치하는 것이 3,662개, 불일치 하는 것이 8,781개 였다.

기준	원어	우리말샘	IPA
우리말샘과 ipa의 발음 비매칭	abstraction	앱스트랙션	애브스트래크시언
	amoxicillin	아목사실린	어모크서시린
	backhand	백핸드	배크핸드
우리말샘과 ipa의 발음 매칭	detail	디테일	디테일
	fingerprinting	핑거프린팅	핑거프린팅
	guitarist	기타리스트	기타리스트

A DB 와 우리말샘 DB에 동시에 존재하는 단어 예시

IPA DB에 저장된 단어의 양은 우리말샘 DB보다 훨씬 많지만, IPA를 한국어 발음으로 변환하는 규칙이 불완전하기 때문에, 위 표에서 보는 것과 같이, 부정확한 발음도 다수 존재하였다. 따라서 우리의 최종 시스템이 영단어를 한국어 발음으로 변환하는 과정에서 표준화된 우리말샘 DB 를 1순위로 체크하고, IPA DB를 2순위로 사용하는 것이 적절하다고 판단하였다.

3.2. 초기 딥러닝 모델 구성

앞서 언급한 우리말샘 DB와 IPA DB에 존재하지 않는 단어들은 딥러닝 기술을 활용하여 적절한 발음으로 변환하고자 하였다. 먼저 LSTM을 기반으로 한 seq2seq 모델을 구성

하여 학습을 진행하였다.

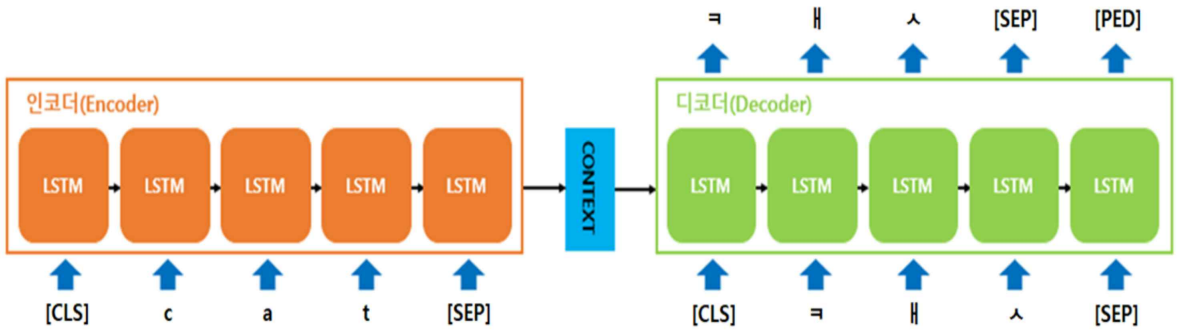


그림 3 구성한 LSTM 모델을 도식화한 모습

학습 데이터로는 우리말샘 DB에 등록된 한국어 발음을 자모 단위로 토큰화 한 후, 학습에 사용하였고, train : 80% validation : 10%, test : 10% 로 데이터를 나누어 학습을 진행하였다. 최종적으로 학습이 완료된 모델은 **55% ~ 60%**의 정확도를 기록하였다.

3.2.1 초기 모델의 문제점 및 개선 사항

앞서 완성한 모델에 테스트 문장을 넣어 실제 체감 성능을 테스트 해보았는데, 몇가지 문제점을 확인할 수 있었다.

[ex] "노승일 K 스포츠재단 부장이 보관한 포스트잇 다섯장."

실제 발음: **케이** / 딥러닝 모델 변환 발음 : **케 ㅣ**

위의 예시에서 확인할 수 있듯이, 자음 모음이 단독으로 출현하는 경우가 발생하였다. 이는 학습 데이터를 구성할 때, 자모 단위로 토큰화 하여 학습한 것이 원인으로 추정되

었다.

[ex] "최근에 발전한 수많은 기술 덕분에 IOT가 실용화되었습니다.

실제 발음: **아이오티** / 딥러닝 모델 변환 발음 : **이오트**

또한, 위와 같이 대문자로만 이루어진 단어들을 정확한 발음으로 변환하는데 특히 어려움을 겪었다. 이는 학습 데이터의 대부분이 소문자로만 이루어진 단어이기 때문에 발생한 것으로 예상되었다.

위에서 제시된 문제점들을 개선하기 위해서, 최종 시스템의 변환 과정을 다음과 같이 수정하기로 결정하였다.

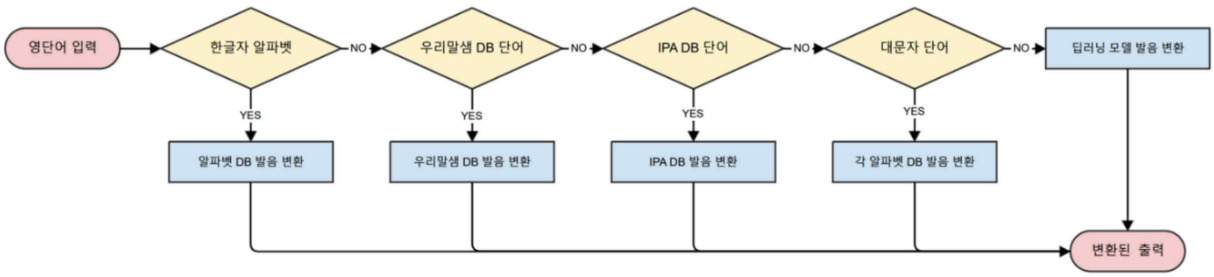


그림 최종적으로 확정된 시스템의 Flowchart

그림7 에서 보는 것과 같이, 우리말샘 DB와 IPA DB를 체크 하는 과정의 앞 뒤에, 새로운 과정을 추가하였다. 여러 문장들을 분석해본 결과, 한 글자 알파벳이나 대문자로만 이루어진 영단어는 대부분의 경우, 알파벳 문자 그대로 발음(EX 비타민c, IOT, WCDMA 등)

하는 것을 알 수 있었다.

그렇기 때문에, a~Z 까지 모든 알파벳이 저장된 알파벳 DB를 구축한 뒤, 위의 경우에 해당 되면 알파벳 DB에 등록된 발음으로 한 글자씩 끊어서 발음하기로 하였다.

한 글자 알파벳과 대문자로만 이루어진 단어를 우리말샘 DB를 체크하기 전, 첫 번째 과정에서 한번에 처리할 수도 있었으나, 'NASA(나사)' 와 같이 대문자로만 이루어졌지만, 알파벳 문자 그대로 발음하지 않는 예외 단어도 일부 존재하였다. 이러한 단어는 대부분 우리말샘 DB와 IPA DB에 등록이 되어 있기 때문에, DB를 이용하여 먼저 정확한 발음으로 변환하고, DB에 등록되어 있지 않으면서 대문자로만 이루어진 단어에 대해서만 알파벳 문자 그대로 발음하기로 한 것이다.

이러한 변경을 통해, 딥러닝 모델로 변환하면 정확도가 매우 떨어지는 단어들을 사전에 필터링할 수 있게 되었다.

3.3. 딥러닝 모델 수정

앞서 생성한 LSTM 모델의 경우 정확도가 55% ~ 60% 였고, 하이퍼 파라미터를 다양한 방식으로 튜닝을 해보았지만, 더 이상 큰 개선의 여지가 보이지 않았다. 이정도 정확도의 딥러닝 모델로는 정확한 발음 변환을 수행할 수 없다고 판단하였기 때문에, LSTM 모델 대신 Transformer 모델을 사용하여 학습을 다시 진행해보기로 하였다.

3.3.1. 학습 데이터 개선

기존에 사용했던, 우리말샘 DB를 좀 더 정제하여, 더 많은 학습 데이터를 모델에 적용하고자 하였다. 우리말샘 DB에는 여러 개의 어절로 이루어진 단어들이 다수 존재하는데, 이를 어절 단위로 분리하여, 각각의 단어를 학습 데이터에 추가하기로 하였다. 그리고 'MultiHeadAttention' 과 같이 여러 어절로 이루어졌지만, 띄어쓰기 없이 쓰는 경우가 실제 문장들에 많이 나타났다. 이러한 단어들에 대한 변환 성능을 높이기 위해 다어절 단어의 공백을 제거해서 학습 데이터에 추가하였다.

다음의 그림은 위 과정을 이해 하기 쉽게 도식화 한 것이다.

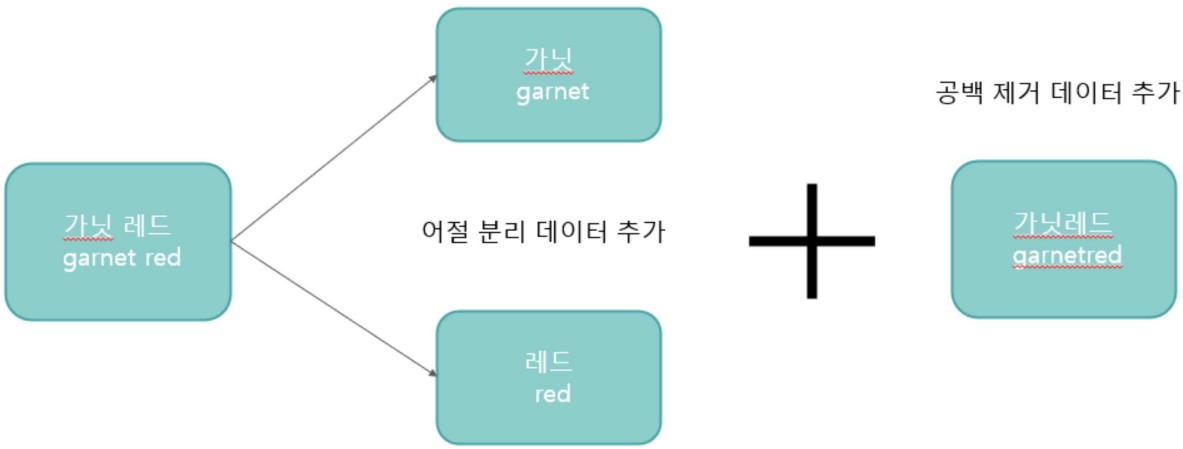


그림 8 학습 데이터 전처리 과정 도식화

위 과정을 통해 전처리한 데이터를 아래의 표와 같이 2가지의 종류로 나누어서 각각에 대해 모델을 생성하도록 하였다.

구분	어절 분리 단어	공백 제거 단어	개수
학습데이터_1(어절 분리 단어를 추가한 우리말샘 DB)	포함	미포함	61,587개
학습데이터_2(학습데이터1 + 공백 제거 단어를 추가한 우리말샘 DB)		포함	100,076개

그림 9 전처리 된 데이터의 개수 현황

정제된 학습 데이터를 활용하여 새롭게 구성하는 Transformer 모델의 하이퍼 파라미터는 다음과 같다.

Hyper Parameter	설정 값
Number of layers	3
Hidden dimension	128
feed_forward_dimension	512
Batch size	128
Number of epochs	500
Optimizer	Adam
dropout	0.1

그림 10 하이퍼 파라미터 설정 값

영단어	LSTM model [학습 데이터_1, Accuray 59.81%]	Transformer model_1 [학습 데이터_1, Accuray 70.52%]	Transformer model_2 [학습 데이터_2, Accuray 82.92%]
once	온스	온스	온스
multiheadattention	멀티헤어드텐션	멀티헤데이텐션	멀티헤드어텐션
homerunball	홈머런볼	홈런볼	홈런볼
breakingdrive	브레이킹드드라이ㅂ	브레이킹드라이브	브레이킹드라이브
butterfly	버터플라이	버터플라이	버터플리
multilayersequence	멀티레이어시퀀스	멀티레이어시퀀스	멀티레이어시퀀스
sequencediagram	시퀀스다이어그램	시퀀스다이어그램	시퀀스다이어그램
whitecotton	화이트코튼	화이트코튼	화이트코튼

그림 11 새롭게 구성한 모델과 LSTM의 성능 비교

새로운 학습 데이터를 활용하여, 생성한 Transformer 모델의 정확도는 각각 **70.52%** 와 **82.92%** 를 기록했으며, 정확도가 대략 60% 정도에 그쳤던 LSTM 모델에 비해, **적게는 10% 에서 최대 20% 가량 정확도가 향상된** 것을 확인할 수 있었다.

4. 연구 결과 분석 및 평가

4.1. 성능 평가에 사용된 테스트 데이터 설명

최종적으로 완성된 시스템의 성능을 평가 하기 위해서 AI Hub 의 데이터를 이용하였다. AI Hub 란, AI 기술 및 제품, 서비스 개발에 필요한 AI 데이터를 지원함으로써 어느 누구나 활용하고 참여 가능한 AI 통합 플랫폼이다.

결과 평가 및 분석을 위해 AI Hub에서 제공하는 음성 합성 학습용 데이터 중에 테스트 데이터로 활용 가능한 데이터들을 수집하여 딥러닝 모델 및 최종 시스템의 성능 평가를 진행하였다.

4.1.1. 테스트 데이터 분류 및 통계

아래는 시스템 성능 평가에 사용 가능한 데이터를 표로 정리한 것이다.

	단어 개수	전체 대비 비율
한 글자 알파벳	1개	0.68%
우리말샘 DB 내 존재	98개	66.2%
IPA DB 내 존재	37개	25%
대문자로만 이루어진 단어	0개	0%
딥러닝 처리	12개	8.1%
총 단어	148개	100%

그림 12 AI Hub 데이터 통계

최종 시스템에 AI Hub 데이터를 넣어 본 결과, 한 글자 알파벳은 1개가 있었고, 우리말샘 DB 내 존재하는 단어는 총 98개로 과반수 이상을 차지하였다. 그 이외의 단어들 중 IPA DB에서 37개의 단어들을 변환할 수 있었고, 대문자로만 이루어진 단어는 없었다. 최종적으로 딥러닝으로 처리할 수 밖에 없는 단어는 총 12개가 존재했다.

4.2. 딥러닝 모델 분석 및 평가

4.2.1. 단일 모델의 성능 분석

AI Hub 데이터 중 딥러닝으로 처리할 수 밖에 없는 단어 12개에 대해서, 앞서 구축한 Transformer 모델 두 가지를 활용하여 정확도를 테스트 해 보았다.

영단어	Transformer model_1[Accuray 75%]	Transformer model_2[Accuray 41.7%]
pollicis	폴리시스	폴리시
longus	롱서스	롱스
carpi	카피	카피
radialis	라디알리스	라디알
gervaise	거베이스	거비즈
paleface	페일페이스	팰리피스
communitybased	커뮤니티베이스트	커뮤니티베이스드
remitting	리미팅	리미팅
khyber	키버	키버
untamed	언타메드	언탐
apicius	아피슈즈	아피

그림 13 단일 모델 딥러닝 변환 결과

총 12개의 단어 중 중복을 제외한 11개의 단어를 딥러닝 모델을 이용하여 한국어 발음으로 변환 해 보았다. Transformer Model_2가 더 높은 정확도를 보일 것이라고 예상하였지만, Model_1이 75% 로 더 높은 성능을 보였다. 11개의 단어만으로는 모델의 정확한 성능을 평가하기에 어렵다고 판단되어, 최종 시스템에서 우리말샘 DB로 변환했던 98개의 단어와 IPA DB로 변환한 37개의 단어를 합친 135개의 단어를 활용하여 추가 테스트를 진행하였다.

	총 단어	정답 개수	정확도
Transformer model_1	135	68	50.4%
Transformer model_2	135	80	59.3%

그림 14 모델 별 변환 정확도

그림 13에서 Model_1 의 정확도가 더 높게 나왔던 것과는 다르게, 그림 14의 결과에서는 Model_2의 정확도가 더 높게 나왔다. 이러한 현상의 원인을 분석하기 위해, 테스트에 사용되었던 135개의 단어들의 변환 결과를 보다 면밀히 분석 해 보았다. Model_1의 경우, 'the', 'son', 'end' 와 같이 길이가 매우 짧은 단어들은 각각 '더더더더', '손넌', '엔드덴드' 로 잘못 변환하는 것을 볼 수 있었다. 하지만 10글자 내외의 단어들에 대해서는 Model_2 보다 변환 성능이 오히려 높은 것을 알 수 있었다.

Model_2의 경우에는 위에서 언급한 길이가 매우 짧은 단어들을 각각 '더', '손', '엔드' 로 적절하게 변환 해 주었고, 조금 더 길이가 길고, 복잡한 단어들에 대해서 Model_1 보다 정확하게 예측 해 주는 것을 알 수 있었는데 'documentary' 나 'multi head attention' 과 같은 단어들이 그 예이다.

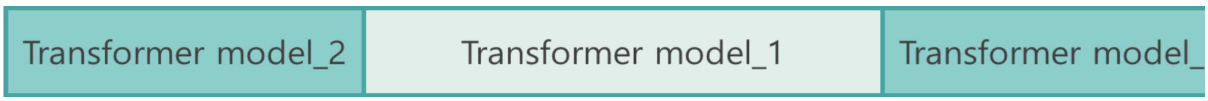
위 결과에서 알 수 있듯이, 각각의 모델 별로 강점이 존재하므로, 두 모델을 적절한 방식으로 결합하면 더 좋은 결과를 예측해 낼 것이라고 추측할 수 있었고, 적절한 기준을 세워 모델을 결합 해 본 뒤, 테스트를 다시 진행하기로 하였다.

4.2.2. 두 모델의 통합 기준 및 성능

아래의 표에서 볼 수 있듯이, Model_1 은 짧은 길이의 단어와 중간 길이의 단어에 대해서 변환 성능이 높았고, Model_2 는 매우 짧은 단어나, 길이가 길고 복잡한 단어에 대해 변환 성능이 높았기 때문에, 아래 그림과 같이 기준을 나누어 모델을 결합 해 보기로 하였다.

단어	길이	Transformer model_1	Transformer model_2
we	2글자	위웨	위
unsqueeze	9글자	언스퀴즈	언스키즈
informationsecurity	19글자	인포메이션스큐리티	인포메이션시큐리티

그림 15 입력 길이 차이에 따른 모델별 결과



4글자 12글자

그림 16 Transformer model_1+2 구성도

먼저 최종 시스템에 입력했을 때, 딥러닝으로 처리되는 단어(그림13)에 대해서 통합 모델의 성능 테스트를 진행 해 보았다.

영단어	Transformer model_1[Accuray 75%]	Transformer model_2[Accuray 41.7%]	Transformer model_1 + 2[Accuray 83.3%]
pollicis	폴리시스	폴리시	폴리시스
longus	롱서스	롱스	롱서스
carpi	카피	카피	카피
radialis	라디알리스	라디알	라디알리스
gervaise	거베이스	거비즈	거베이스
paleface	페일페이스	팰리피스	페일페이스
communitybased	커뮤니티베이스트	커뮤니티베이스드	커뮤니티베이스드
remitting	리미팅	리미팅	리미팅
khyber	키버	키버	키버
untamed	언타메드	언탐	언타메드
apicius	아피슈즈	아피	아피슈즈

그림 17 단일 모델과 통합 모델의 정확도 비교

그림에서 알 수 있듯이, Model_1 과 Model_2 를 결합한 Model_1 + 2 의 경우, 단일 모델만을 사용했을 때보다 향상된 83.3%의 정확도를 기록하였다.

또한, 우리말샘 DB와 IPA DB에서 변환되었던 135개의 단어(그림14)에 대해서 테스트를 진행한 결과, 아래의 그림과 같이 Model_1 + 2 이 가장 우수한 성능을 보이는 것을 확인할 수 있었다.

	총 단어	정답 개수	정확도
Transformer model_1	135	68	50.4%
Transformer model_2	135	80	59.3%
Transformer model_1 + 2	135	90	66.7%

그림 18 단일 모델과 통합 모델의 정확도 비교2

4.3. 최종 시스템 평가 및 분석

최종 시스템 성능 평가의 신뢰도를 높이기 위해 AI hub 에서 제공하는 또 다른 데이터를 사용하여 실험을 진행하기로 했다. 이번 실험에 사용된 데이터의 개수와 통계는 아래

	단어 개수	전체 대비 비율
한 글자 알파벳	488개	1.8%
우리말샘 DB 내 존재	18915개	71.7%
IPA DB 내 존재	4175개	15.8%
대문자로만 이루어진 단어	746개	2.8%
딥러닝 처리	2065개	7.8%
총 단어	26389개	

그림 19 AI hub 데이터 통계

와 같다.

여기서 사용된 데이터는 실제 텍스트가 아닌 음성 인식 결과를 기록한 데이터이기에, 같은 단어라도 발음이 다르거나, 표준 발음과 일치하지 않게 기록된 경우가 다수 존재하였다. 그래서 'CCTV'와 같은 단어는 표준 발음이 '시시티브이' 이지만, AI hub 데이터는 '시시티비' 와 같은 형태로 저장되어 있어, 우리가 구축한 시스템이 올바른 발음으로 변환했지만, 오답으로 기록되는 경우가 있었다.

그렇기에 일반적인 정확도 측정 방식을 사용할 수 없다고 판단했고, 우리는 정확도 측정을 위해 Edit distance 방식을 사용하기로 하였다. Edit distance 방식은 두 문자열 사이의 유사성을 측정하는 방법 중 하나이다. Edit distance는 한 문자열을 다른 문자열로 변환하는 데 필요한 최소한의 단일 문자 편집 연산(삽입, 삭제, 치환) 수로 정의된다.

예를 들어 정답이 '어드벤스드'일 때 '어드밴스드'는 ㅐ가 ㅔ로 1번의 치환 과정을 거치면 되므로 편집 거리는 1이다.

시스템이 예측한 결과와 정답 발음 간의 편집 거리가 1 이하일 경우, 우리는 이것을 정답으로 간주하기로 하였다.

	edit distance 기준	총 단어	정답 개수	정확도
최종 시스템	1 이하	26,389	20,206	76.6%
	0	26,389	17,379	65.9%

그림 20 최종 시스템 성능

위와 같이, 총 26,389개의 단어 중 edit distance 방식을 적용 할 경우, 20,206개의 단어를 정확히 변환하여, 정확도 **76.6%**를 기록하였다. 이는 대부분의 단어를 정확한 발음으로 변환하는 최종 시스템을 성공적으로 구축하였음을 의미한다.

5. 결론 및 향후 연구 방향

우리는 일반적인 TTS 시스템이 문자를 발음으로 변환할 때, 딥러닝 기술을 중점적으로 활용하는 것과는 달리, 지식베이스와 딥러닝 방식을 결합하여 영단어를 한국어 발음으로 변환하는 시스템을 구축하고자 하였다. 그 과정에서 우리말샘 사전과 IPA 사전을 전처리하여 DB를 구축하였고, DB로 변환이 불가능한 단어 들에 대해 Tranformer 모델을 활용하여 총 5단계에 걸쳐, 영단어를 한국어 발음으로 변환하는 시스템을 구축하였다. 그 결과, 상당 수의 단어를 정확한 발음으로 변환하는데 성공하였다.

IPA 사전을 활용하는 과정에서 IPA 기호를 한국어 발음으로 변환하는 과정이 필요했고, 이 과정에서 부정확한 발음으로 변환된 단어들이 상당 수 IPA DB에 저장되었다. 그래서 최종 시스템의 정확도를 저하시키는 가장 큰 원인은 IPA DB가 되었다. IPA 기호를 한국어로 변환하는 규칙을 좀 더 고도화 한다면, 최종 시스템의 성능을 큰 폭으로 향상시킬 수 있을 것이다.

또한 한국어 고유 명사를 영문화 했을 때, 일반적인 영어 단어와의 발음은 서로 차이가 존재할 수 밖에 없다.

단어	구분	발음	예시
yak	한국어 고유 명사 발음	[약]	yakgwa(약과), Yaksu-dong(약수동)
	영어식 발음	[야크]	Yak(동물- 야크)

그림 21 한국어 고유 명사의 영문화 발음과 일반 영어식 발음의 차이

위의 표에서 '약과'라는 단어가 'yakgwa' 라는 영어로 들어온다면, 시스템은 '야크과' 라는 단어를 예측해 낸다. 현재 우리의 연구에서는 이러한 고유 명사는 입력으로 들어오지 않는다고 가정하고 수행하였기에, 향후 연구에서는 한국어 고유 명사 데이터를 추가적으로 수집하여 지식베이스와 딥러닝 학습 데이터에 사용할 수 있다면 보다 높은 성능의 시스템을 구축할 수 있을 것으로 예상된다.

6. 참고 문헌

[1] Vaswani, A. , Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information ProcessingSystems (pp. 5998-6008)

[2] Kcrong (2017, Nov 15). 영어 단어를 한글로. 머신러닝으로 음역하기. Blog4Study. [Online]. Available: https://blog.devkcr.org/275

[3] 국립 국어원 (불분명한 연도). IPA 표기 세칙. 국립 국어원 공식 웹사이트. [Online]. Available: https://www.korean.go.kr/front/page/pageView.do?page_id=P000124&mn_id=97.

[4] Sequence to Sequence Learning with Neural Networks. Available: https://arxiv.org/abs/1409.3215

[4] Pytorch-transformer-kor-eng. Available: https://github.com/Huffon/pytorch-transformer-kor-eng