# CONTENTS

# 1. 기존 TTS의 문제점과 개선 방안


TTS는 "Text-to-Speech"의 약어로 텍스트를 음성으로 변환하는 기술을 의미함.

대중적으로 사용되는 TTS가 외래어 발음을 얼마나 정확하게 변환하는지 확인하기 위해 테스트 문장을 활용하여 실험함.

papago

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

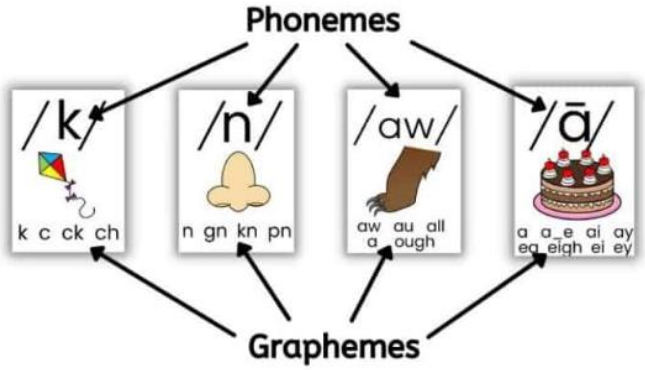
bing

"궉 박사는 현재 유체역학 학술지인 유체물리학지의 편집장과 미국 UCLA석좌교수로 활동하고 있다."
실제 발음: 유시엘에이 / TTS 변환 발음 : 유시알에이

상용화 된 기존 TTS는 밑줄 친 단어를 정확한 한국어 발음으로 변환하지 못함.

# 1. 기존 TTS의 문제점과 개선 방안

G2P 모델의 정확도가 낮은 기존 TTS는 외래어가 포함된 문장을 정확한 한국어 발음으로 변환하지 못함.




G2P란 "Grapheme-to-Phoneme "의 약어로 입력으로 들어온 단어의 철자(표기)로부터 해당하는 발음을 매칭해주는 기술.

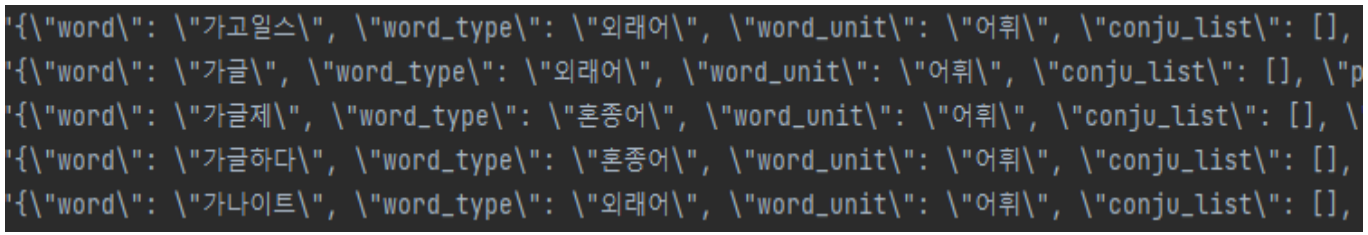
이를 보완하기 위해 우리말샘 사전을 기반으로 하는 지식 베이스 방식의 발음열 기분석 사전을 구축.

# 2. 데이터 전처리

우리말샘 사전의 전체 단어 데이터들을 다운 받아 필요한 형태로 전처리 함.

Sense number 가 001인 경우(사전에서의 첫번째로 뜻으로 사용되는 단어)만 사용.

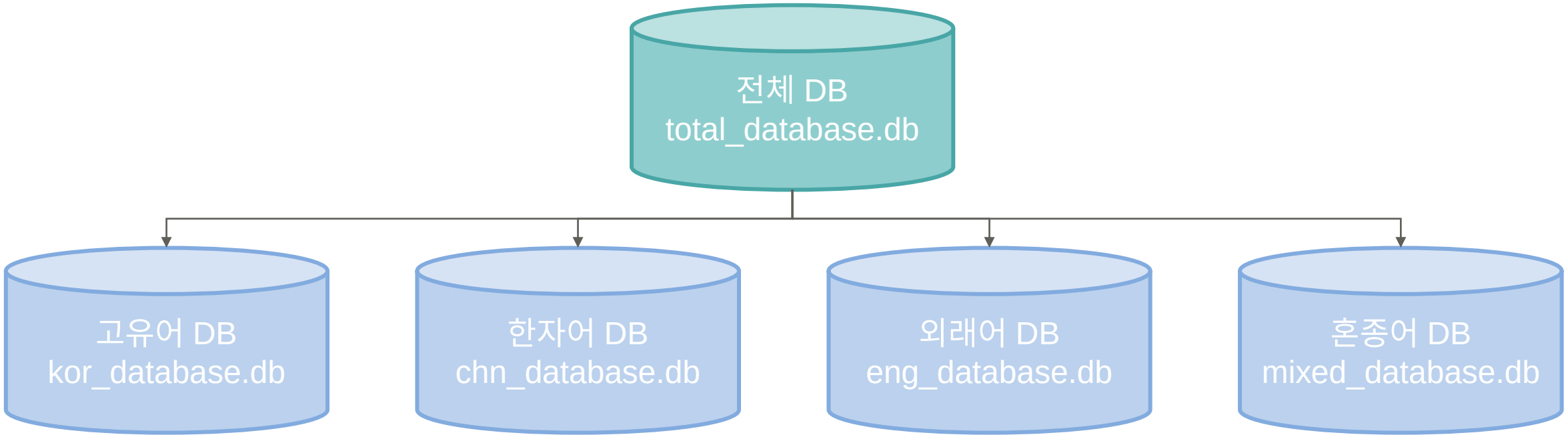
북한어나 방언을 제외한 일반어만 사용.

# 2. 데이터 전처리

우리말샘 전체 데이터를 전처리 과정을 통하여 4가지로 분류.

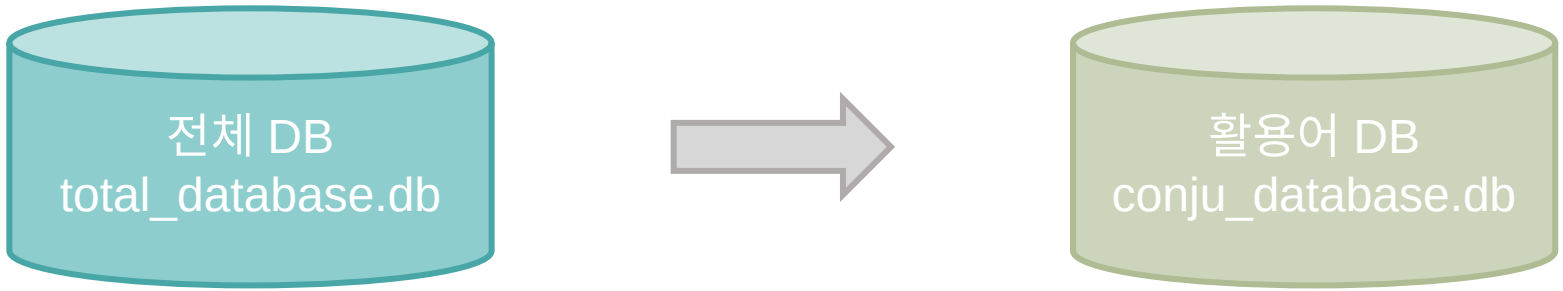
최종적으로 전체를 포함하여 고유어, 한자어, 외래어, 혼종어 5가지 DB로 분류.



# 2. 데이터 전처리

전처리 과정을 거친 전체 DB인 total_database.db에서 활용어의 표기법과 발음법을 추출.

이전에 분류된 5가지 DB와 별개로 활용어 DB인 conju_database.db 생성.



# 2. 데이터 전처리

외래어 DB 속 저장된 데이터 형식 예시

| 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 |

# 3. 데이터 통계

총 24개의 우리말샘의 json 파일

전체 1,164,952개의 단어

| 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 |

전처리 과정을 통해 분류된 DB 파일

전체 779,428개의 단어

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

# 3. 데이터 통계

KT에서 제공한 외래어가 포함된 학습 데이터를 현재 구축한 DB를 이용하여 매칭 시켜보았음.

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

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

※ 대소문자를 구분하여 통계를 낸 이유 : 영어 단어를 DB와 매칭하는 순서가 아래와 같은 과정으로 진행되기 때문.

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.

# 4. 데이터 처리 결과 및 개선점

우리말샘 사전으로 DB를 구축할 때 sense number가 001인 데이터만을 활용하였기 때문에

sense number가 001이 아닐 경우 DB에 존재하지 않아 발음열 매칭이 불가능한 문제가 발생.

Ex) “자석은 N극과 N극이 만나면 서로 밀어냅니다.”

엔(en[円])「001」일본의 화폐 단위. 기호는 ￥.

엔(N / n)「003」영어 알파벳의 열네 번째 자모 이름.

# 4. 데이터 처리 결과 및 개선점

우리말샘으로 분류된 DB에 존재하지 않으면서 대문자인 경우 알파벳을 그대로 발음할 수 있지만

DB에 존재하지 않고 대문자로만 이루어지지 않은 경우 문제가 발생.

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

따라서 이러한 외래어를 문제 없이 한국어로 발음 변환을 하기 위해 딥러닝 모델 구축이 필요함.

# 5. 향후 진행 계획

우리말샘 사전을 기반으로 최대한 많은 외래어를 DB에 저장 후 활용하려고 했으나

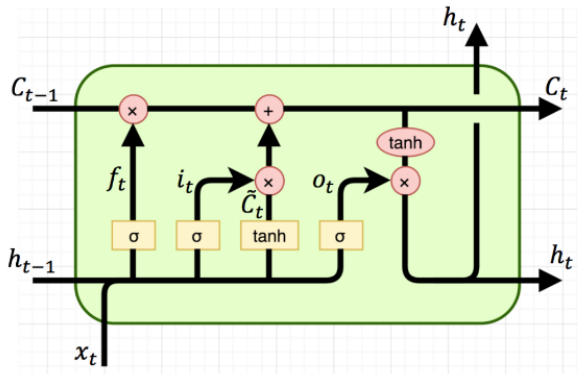
sense number로 인해 DB에 저장되지 않은 단어와 우리말샘 사전에 미등재된 단어로 인해

DB에 존재하지 않는 단어들이 다수 존재.

이를 보완하기 위해 Transformer 모델 혹은

LSTM 모델을 활용한 딥러닝 모델을 구축하여

DB에 존재하지 않는 단어를 처리할 계획임.




(a) Long Short-Term Memory

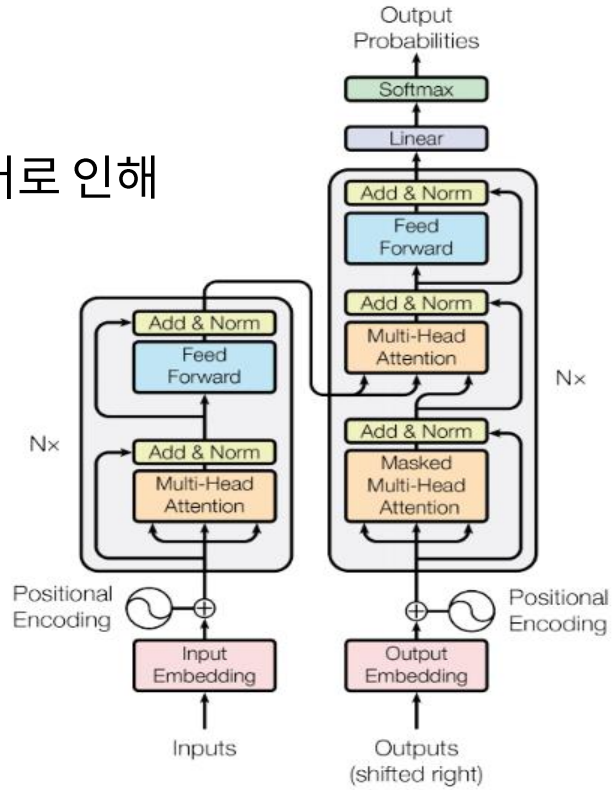



Figure 1: The Transformer - model architecture.