한국어 말하기 능력 시험 (TOPIK)

AI 학습 도우미

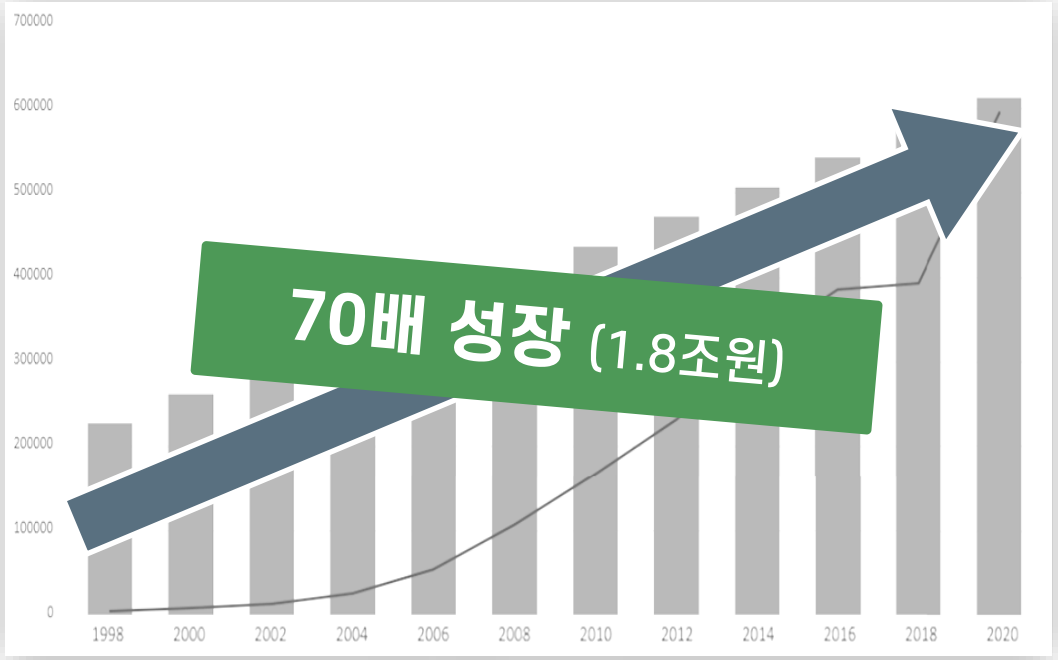
인공지능사관학교 3기

AI 모델링 - 언어지능

KoPanda

성장하는 시장, 부족한 인프라

매년 증가하는 응시자

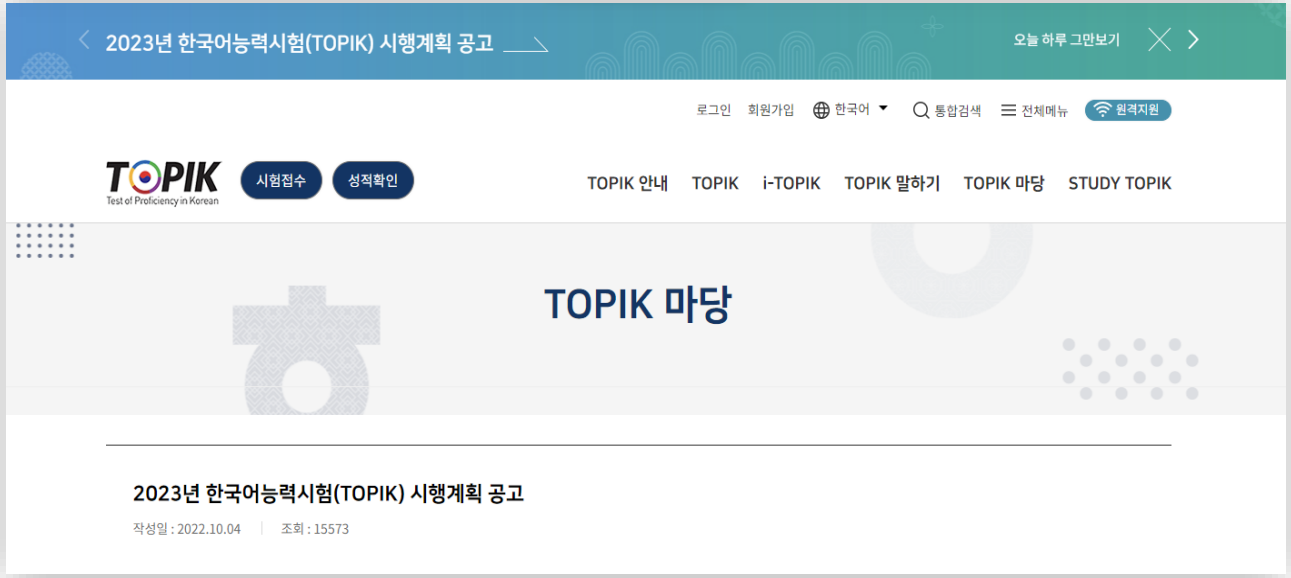


(교육부) 연도별 한국어능력시험(TOPIK) 응시자 현황 (1997 ~ 2021)

치열한 시험 접속

(KBS 뉴스) 2021년 10월 09일 / TOPIK 공식 사이트

TOPIK 말하기 시험의 도입



학습자의 요구

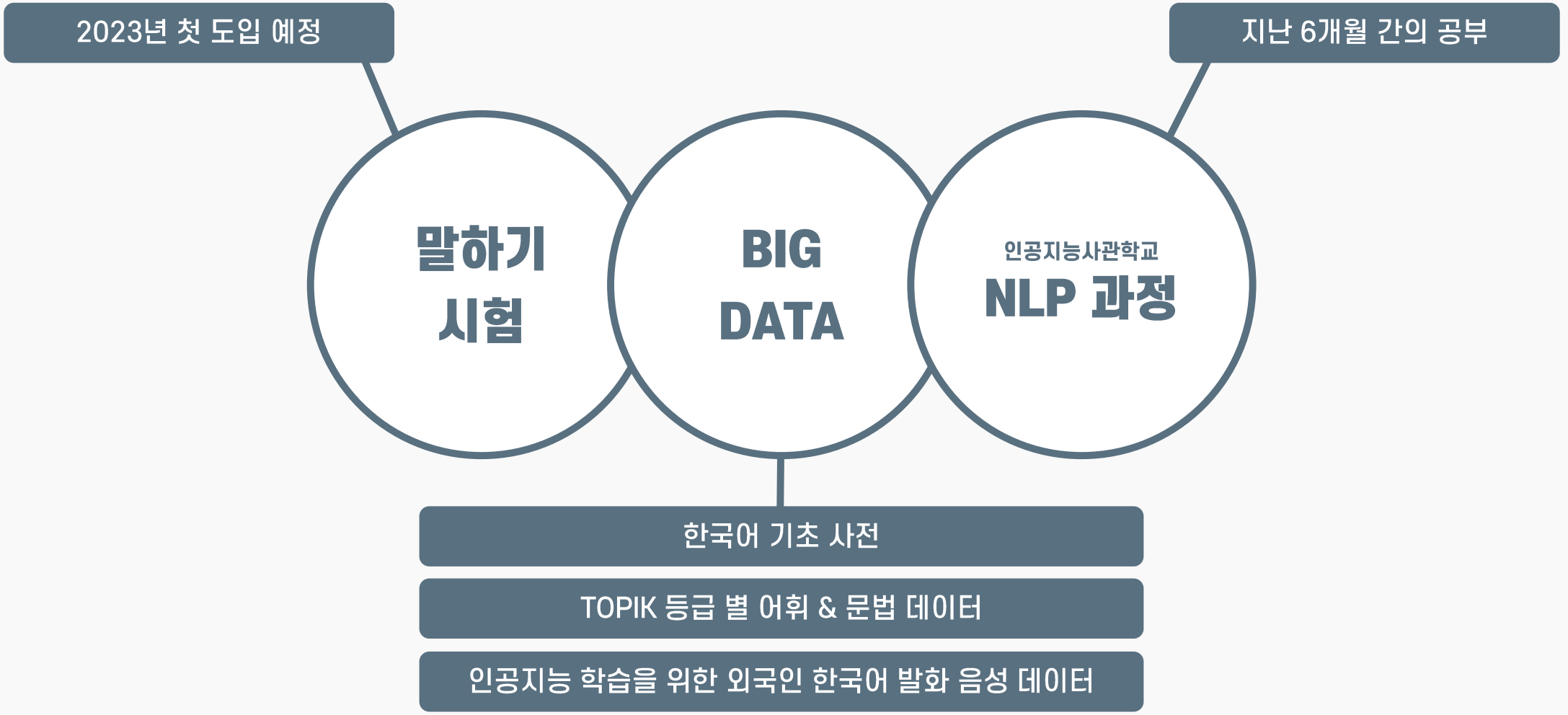
학습자

TOPIK 말하기 모의 시험

평가 & 피드백

보충 학습

최적의 개발 타이밍

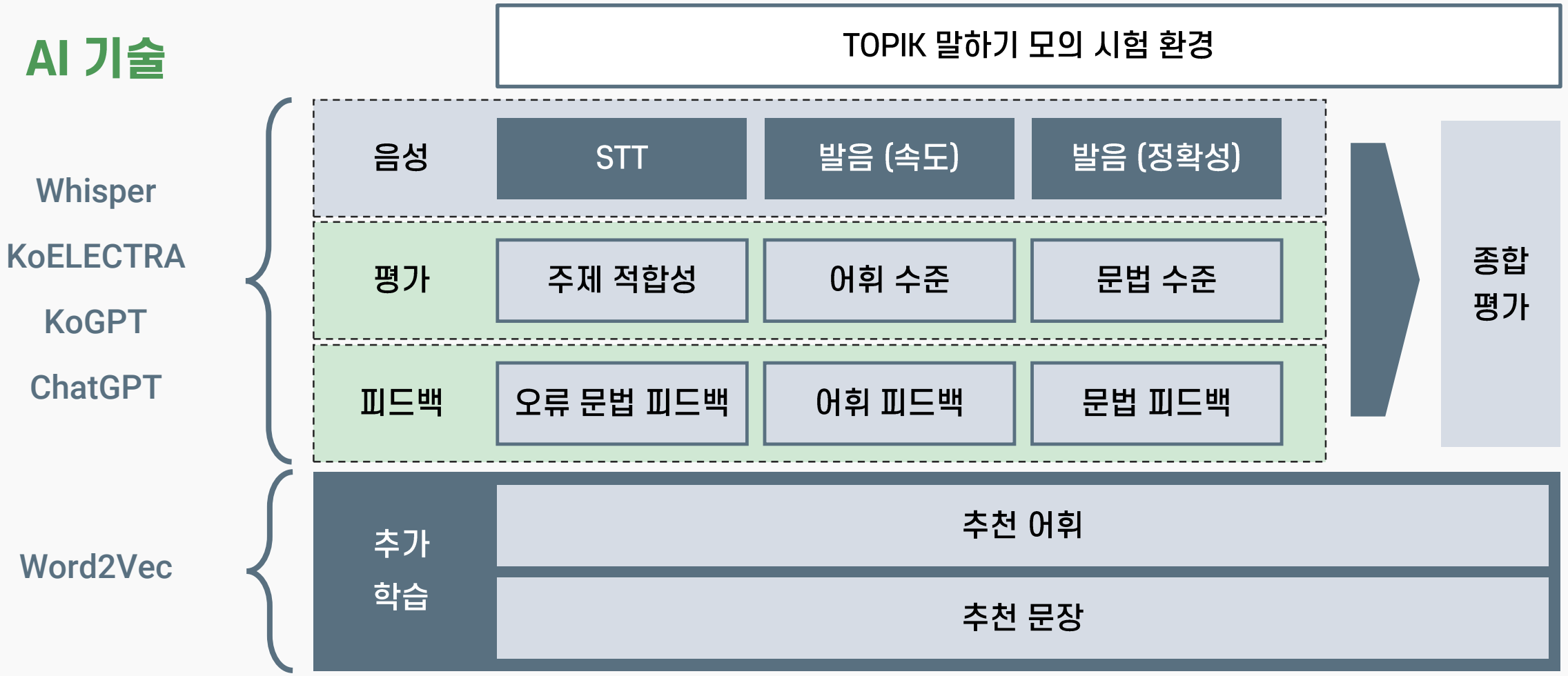

2023년 첫 도입 예정
지난 6개월 간의 공부
말하기 시험
BIG DATA
인공지능사관학교
NLP 과정
한국어 기초 사전
TOPIK 등급 별 어휘 & 문법 데이터
인공지능 학습을 위한 외국인 한국어 발화 음성 데이터

한국어 말하기 능력 시험 (TOPIK)

AI 학습 도우미

기존 방법 vs KoPanda

KoPanda 구성


AI 기술
TOPIK 말하기 모의 시험 환경
Whisper
KoELECTRA
KoGPT
ChatGPT
음성
STT
발음 (속도)
발음 (정확성)
평가
주제 적합성
어휘 수준
문법 수준
피드백
오류 문법 피드백
어휘 피드백
문법 피드백
종합 평가
Word2Vec
추가 학습
추천 어휘
추천 문장

KoPanda 알고리즘

음성 → **Web Server** → 텍스트 → **Model Server** → 평가 & 피드백

STT (Speech To Text)

음성을 텍스트로 변환
발음 속도 및 정확도 평가

STT
WER 8 CER 3
Transformer

주제 평가

주제 적합성 평가

주제 적합성
이진 분류 : 0.93
상/중/하 분류 : 0.63

발음

음성의 발음 속도 평가
음성의 발음 정확도 평가

< 발화 속도와 한국어 분절음의 음향학적 특성 >

느린 발화	정상 발화	빠른 발화
4.21 ~ 4.8 음절/초	5.6 ~ 6.29 음절/초	7.04 ~ 8.14 음절/초

어휘

어휘 평가 / 피드백

문법

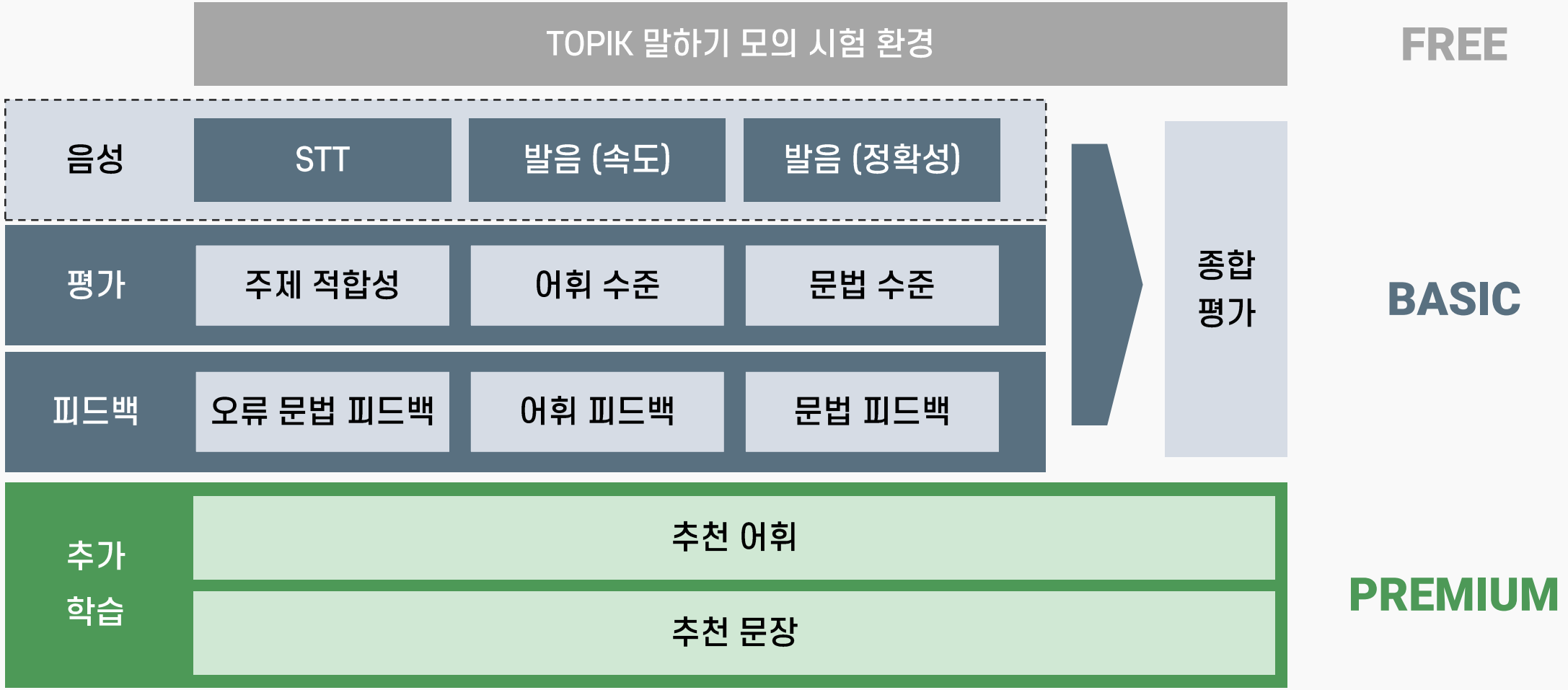
문법 평가 / 피드백
오류 문법 피드백

추가 학습

추천 어휘
추천 문장

시연 영상

KoPanda


TOPIK 말하기 모의 시험 환경
FREE
음성
STT
발음 (속도)
발음 (정확성)
평가
주제 적합성
어휘 수준
문법 수준
피드백
오류 문법 피드백
어휘 피드백
문법 피드백
종합 평가
BASIC
추가 학습
추천 어휘
추천 문장
PREMIUM

KoPanda

TOPIK 말하기 모의 시험 환경

음성	STT	발음 (속도)	발음 (정확성)
평가	주제 적합성	어휘 수준	문법 수준
피드백	오류 문법 피드백	어휘 피드백	문법 피드백
추가 학습	추천 어휘		
	추천 문장		

종합 평가

FREE

매일 무료 주제

9개

KoPanda 모의 시험 환경

1번. 질문을 듣고 대답하십시오.
20초 동안 준비하십시오.
'삐' 소리가 끝나면 30초 동안 말하십시오.

① 문제 제시

음성 녹음이 시작되었습니다.
30초 동안 말하십시오.

② 녹음

코펜더 님의 멋진 답변을 다시 들어보세요!

한글로 옮겨진 나의 답변을 확인해보세요.

③ 답변 다시 듣기

TOPIK 예상 문제

KoPanda가 제작한 TOPIK 말하기 주제별 문제 모음입니다.
랜덤 문제를 하루에 각 주제별 1개씩 TOPIK과 유사한 환경에서 연습하실 수 있습니다.
예시 답안과 KoPanda AI 서비스는 제공되지 않습니다.

KoPanda

매일 무료

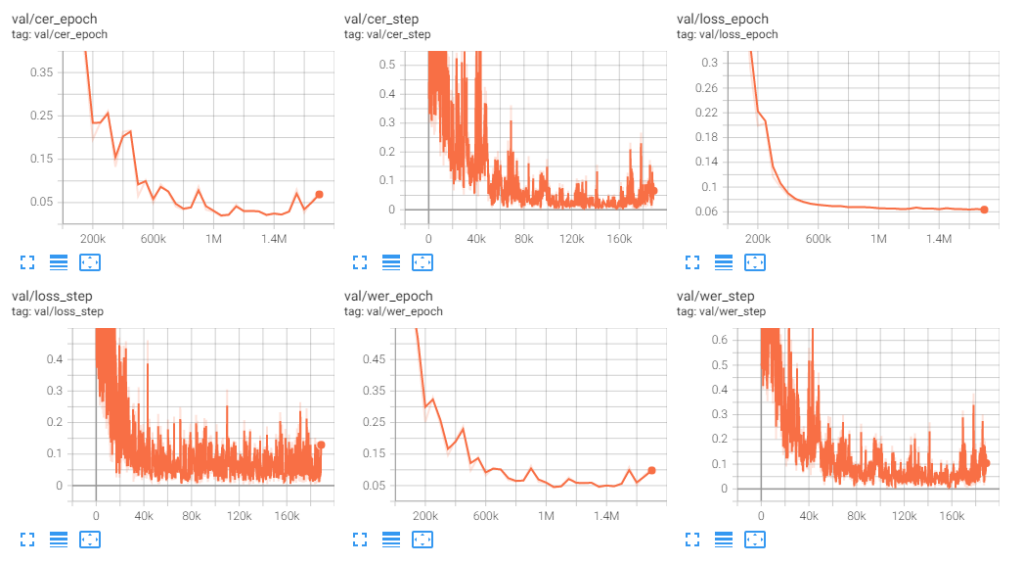
BASIC

1회 평가

1,500원

음성

STT (Speech To Text)



	Kaldi	경쟁업체	Whisper
WER	25	23	**8**
CER	9.5	8	**3**
사용 기술	LSTM, Transformer	LSTM	**Transformer**

* WER (Word Error Rate) : 단어 에러 비율
* CER (Character Error Rate) : 음절 에러 비율

예시 화면

음성

발음 (속도, 정확도)

발음 속도 평가 기준

< 발화 속도와 한국어 분절음의 음향학적 특성 >

느린 발화	정상 발화	빠른 발화
4.21 ~ 4.8 음절/초	5.6 ~ 6.29 음절/초	7.04 ~ 8.14 음절/초

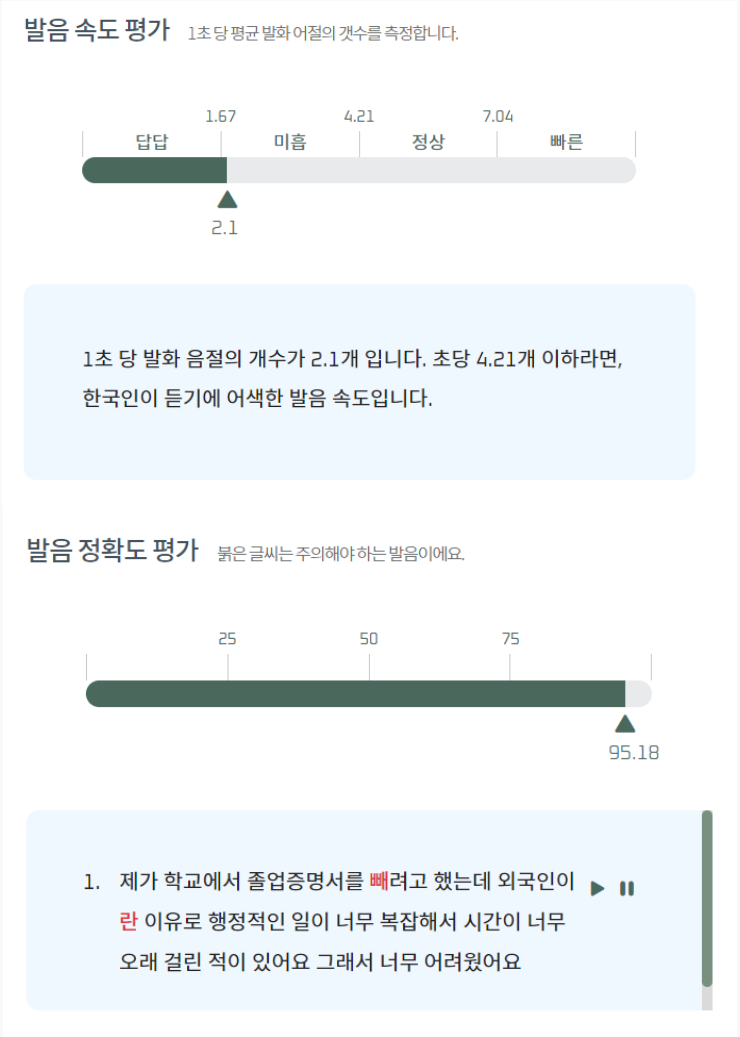
이숙향, 고현주. (2004). 발화속도와 한국어 분절음의 음향학적 특성. 한국음향학회지, 23(2), 162-172.

발음 정확도 평가 기준

- **Whisper (Large model)**
- **Whisper (Small model, Fine Tuning)**

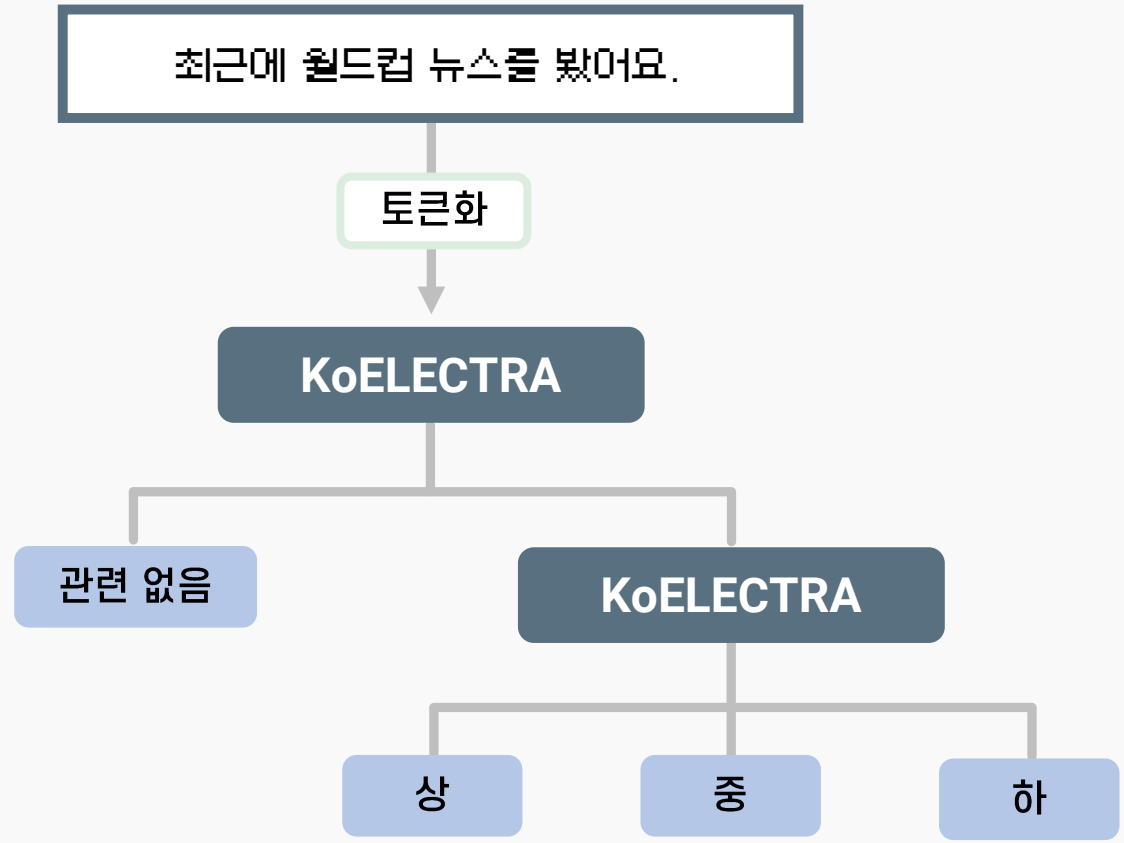
2개의 STT모델(large 모델, 파인튜닝 된 small 모델)을 이용하여 전사된 텍스트 결과를 비교해, 발음 정확도를 구합니다.

예시 화면

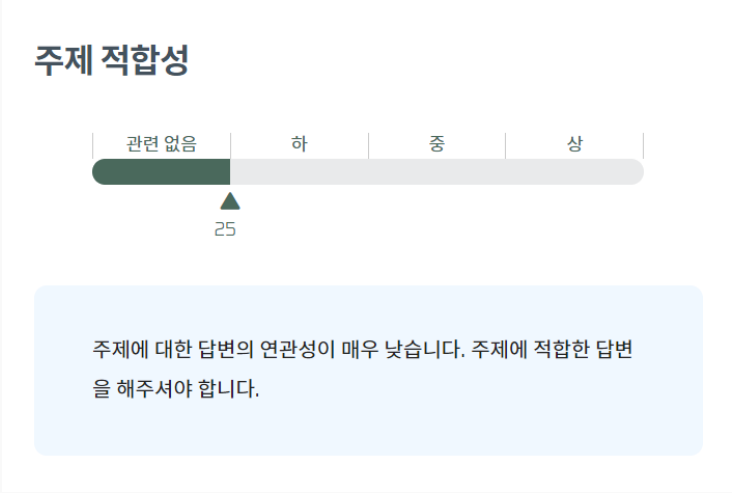


주제 평가

주제 적합성 : 전체 구조

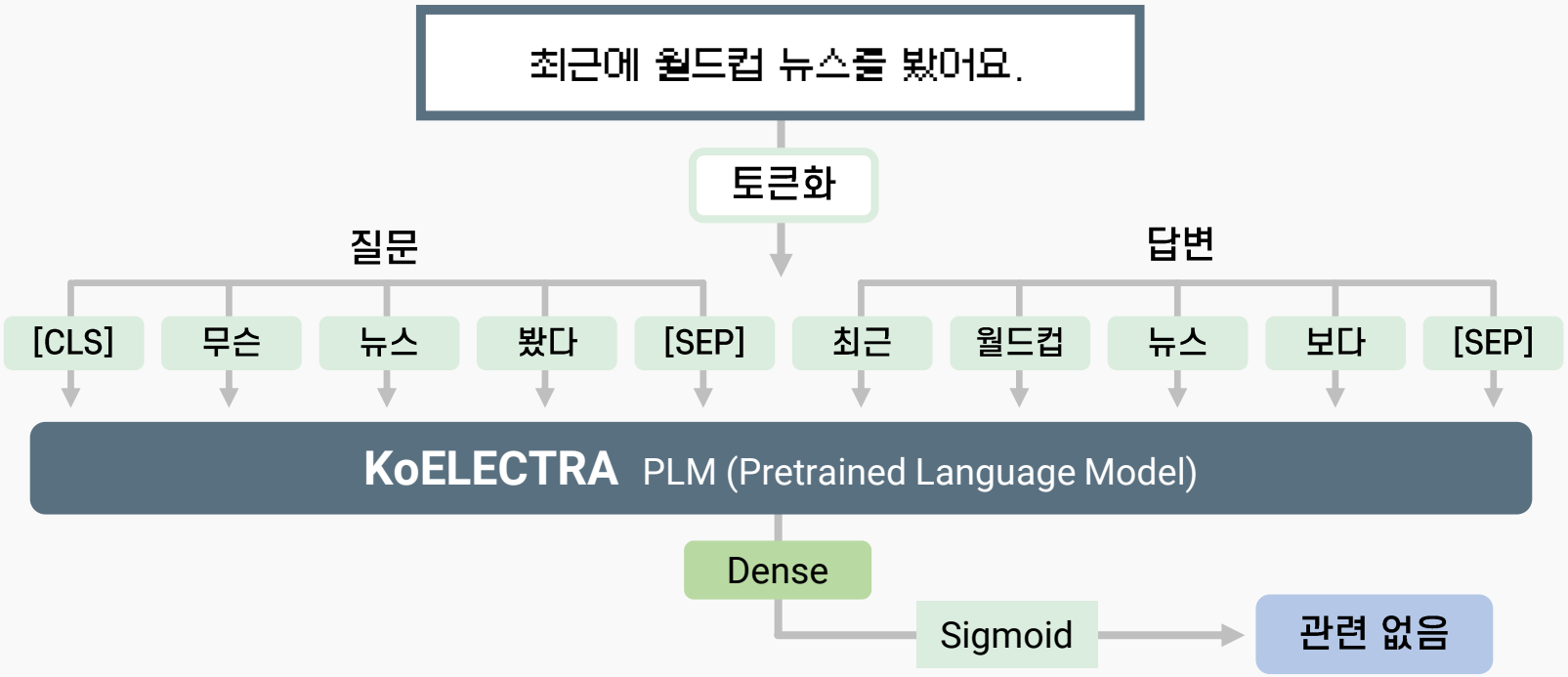


예시 화면



주제 평가

주제 적합성 : 이진 분류



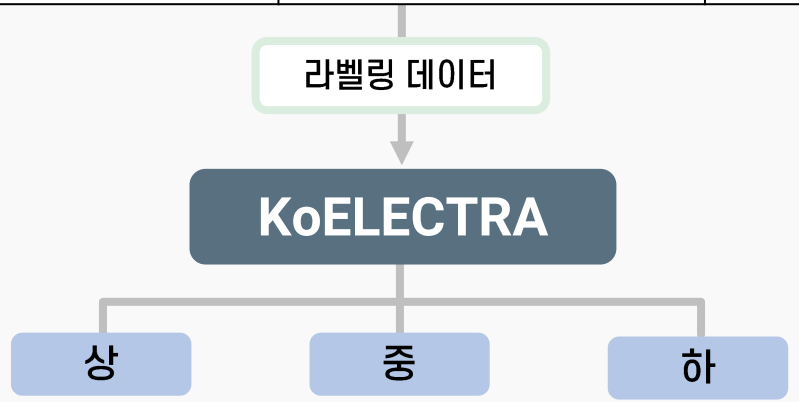
	KakaoBrain / KoGPT	OpenAI / ChatGPT
데이터 크기	375 질문 * 각 20문장 (7,500문장)	15질문 * 각 25문장 (375 문장)
Accuracy	0.93	0.92

예시 화면

주제 평가

주제 적합성 : 상/중/하 분류

	상	중	하
질문 연계성	질문과 연관된 구체적인 답변, 질문 요구사항들이 충실히 반영된 답변	질문과 연관되어 있으나, 질문에서 제시한 예시만 포함되었거나 요구사항이 다 절반 이상만 충족된 문장	질문과 연관 없는 답변, 요구사항을 절반 이하로 포함하는 답변

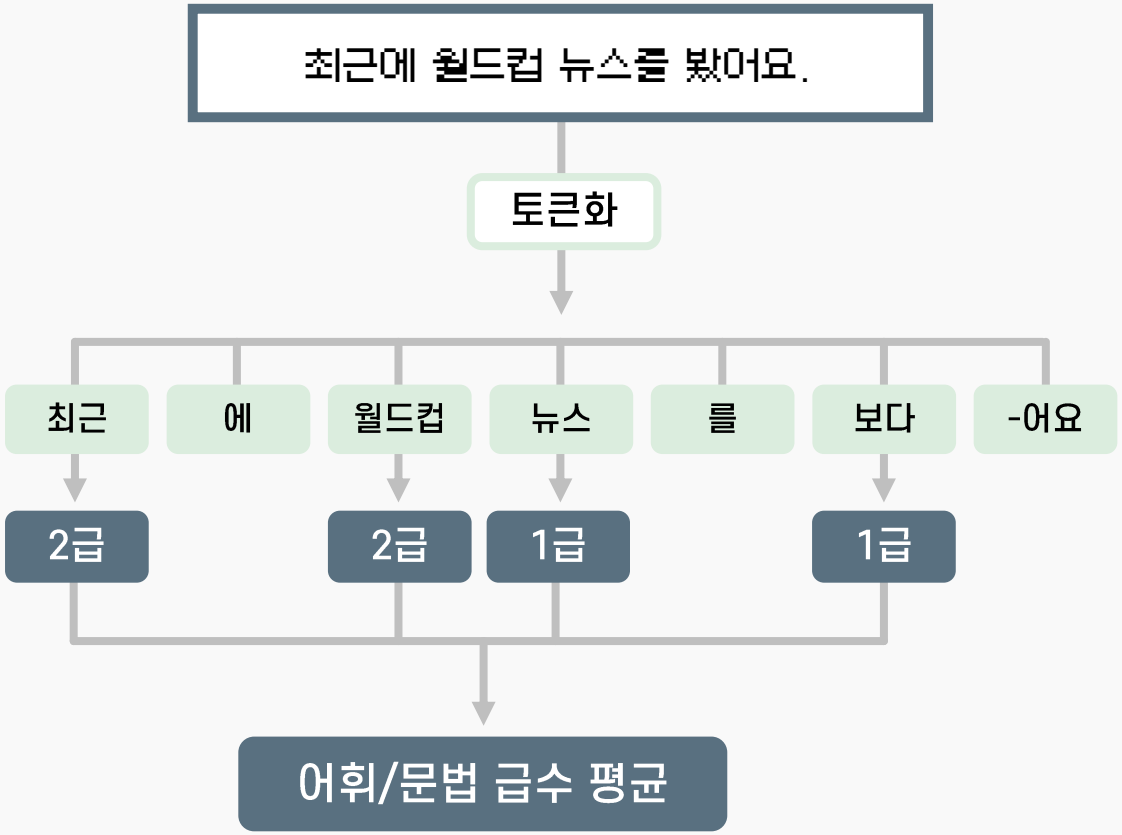


	훈련 세트	검증 세트
데이터 크기	127 질문 * 각 60문장 (7762 문장)	127질문 * 각 10문장 (1270 문장)
F1-Score	-	0.63

예시 화면

어휘, 문법 평가 & 피드백

어휘, 문법



어휘 등급 사전

	어휘_	등급_
0	가	4
1	가게	1
2	가격	1
3	가계	6
4	가곡	5
...	...	...
9989	힘없다	5
9990	힘없이	4
9991	힘입다	6
9992	힘주다	5
9993	힘차다	4

9994 rows × 2 columns

문법 등급 사전

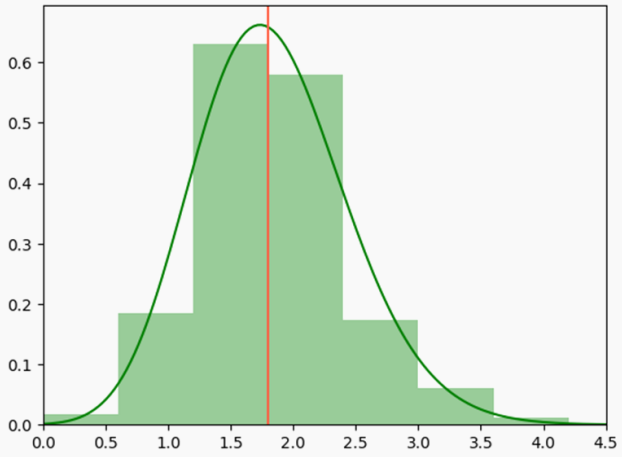
	문법	등급
0	이	초급
1	과	초급
2	까지	초급
3	께서	초급
4	은1	초급
...	...	...
215	-어서인지	중급
216	에 따라	중급
217	에 비하여	중급
218	에 의하여	중급
219	-어 버리다	중급

202 rows × 2 columns

어휘, 문법 평가 & 피드백

어휘, 문법

외국인 한국어 발화 23만 문장을 대상으로 어휘/문법 등급 분포 조사



어휘 백분위 환산 점수

백분위	0	10	20	30	40	50	60	70	80	90
평균단어등급	0.0	1.12	1.33	1.5	1.67	1.8	1.95	2.0	2.25	2.57
점수	10	20	30	40	50	60	70	80	90	100

문법 백분위 환산 점수

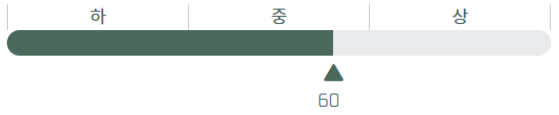
백분위	0	10	20	30	40	50	60	70	80	90
평균문법등급	0.0	0.5	0.75	1.0	1.0	1.2	1.33	1.5	1.7	2.0
점수	10	20	30	40	50	60	70	80	90	100

예시 화면

어휘 등급

어휘	등급	유의어	반의어	길잡이말
너무	1급			너무 심하다
걸리다	1급			그림이 걸리다
적	4급			적을 공격하다

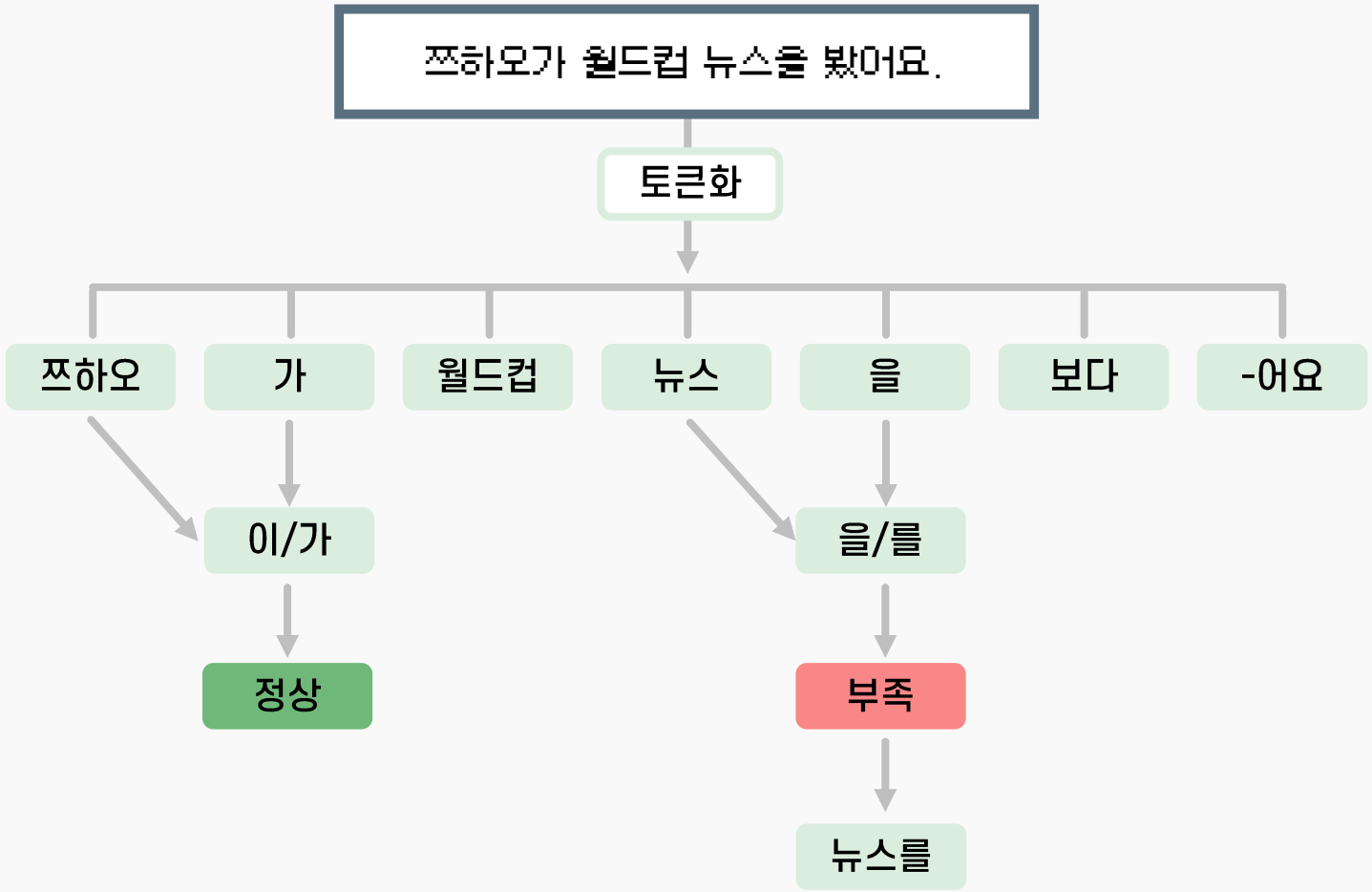
문법 등급



대표형	관련형
이랑	랑
에서	서2
-는데1	-은데1, -ㄴ데1

오류 문법 피드백

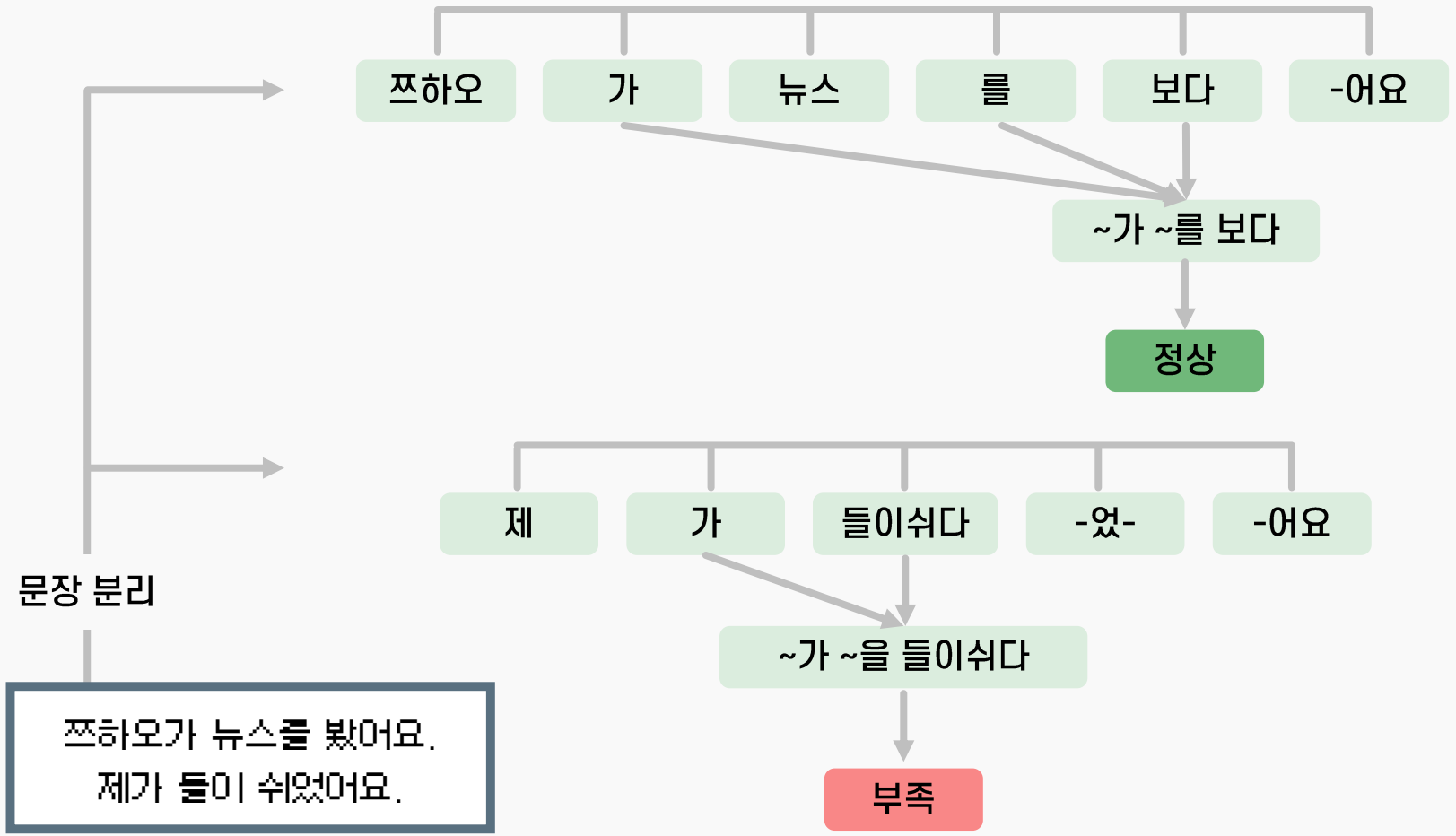
문법 완성도



예시 화면

오류 문법 피드백

문장 완성 형태



예시 화면

종합

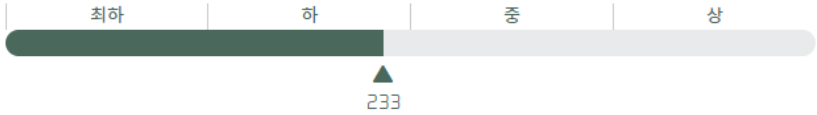
종합 평가

TOPIK 말하기 모의고사

ID	모의고사 회차	응시일
판다스	1회	2022.12.14

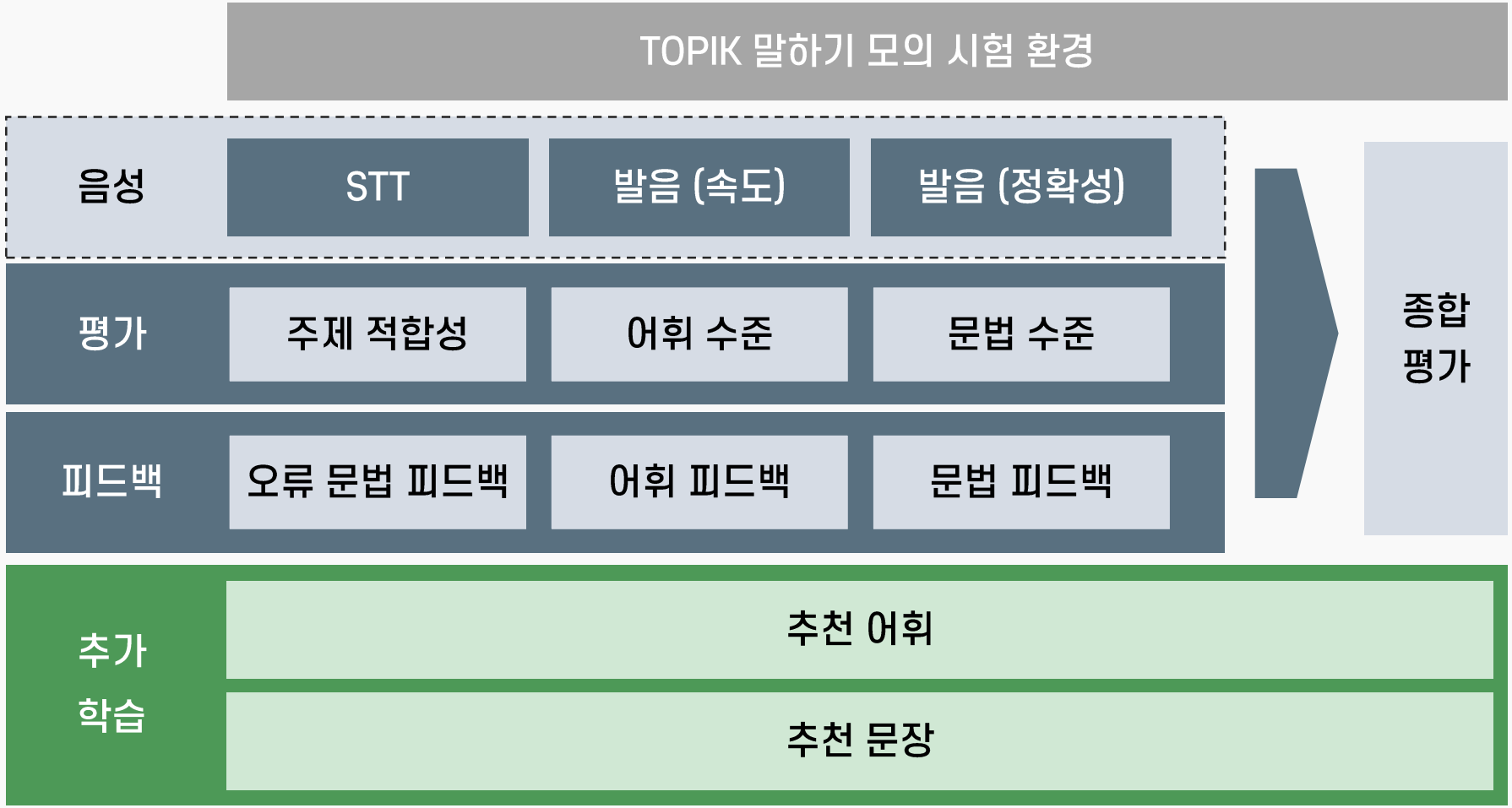
말하기 영역	점수	총점	종합평가
주제 적합성	50/100	233/500	하
발음	84.04/100		
속도	33.0/100		
문법	60/100		
어휘	28/100		

종합 평가



주제에 대한 연관성이 낮은 답변입니다.
질문에 대한 답변 완성도가 부족하고, 어휘와 문법 사용이 부자연스럽습니다.

KoPanda



PREMIUM

월 구독

무제한

9,900원

추가 학습

문장 추천

- 한국에서 쇼핑하기 좋은 장소는 집 근처에 있는 스타 필드입니다 그 이유는 스타 필드는 여러 가게가 모여 있어서 사고 싶은 것이 한 번으로 해결이 됩니다
- 여행을 왔을 때는 마트가 최고예요 한꺼번에 많은 물건을 볼 수도 있고 사람들이 어떻게 생활하는지 느낄 수 있어서요
- 여의도 아이에프씨몰이요 거기 가면 거의 다 구매하고 싶은 거 있고 서점도 있고 먹을 때도 있고 한 번에 다 해결이 돼서요

어휘 피드백 모델

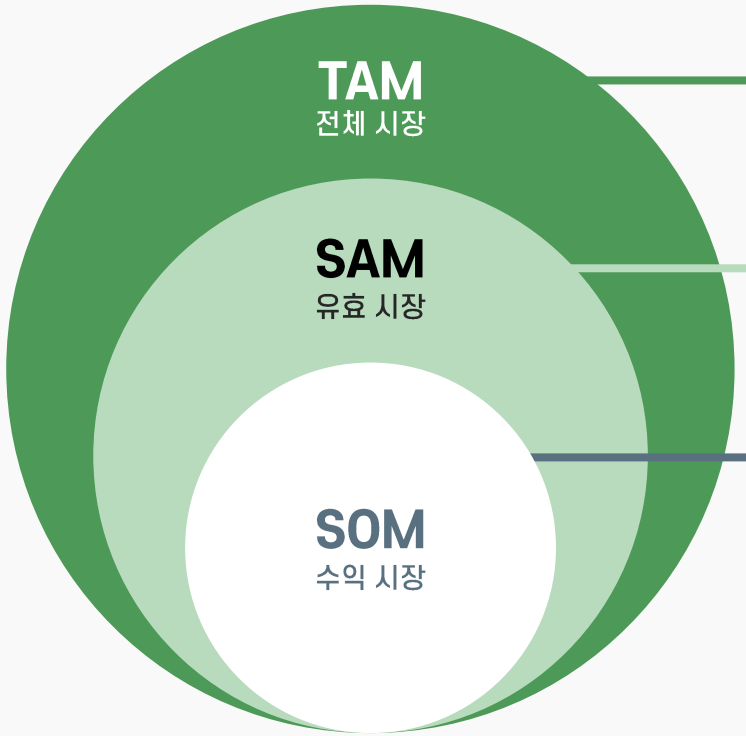
어휘 추천

어휘	중국어	등급	품사	뜻	뜻 (중국어)
때	时，时候	4급	명사	시간의 어떤 순간이나 부분.	指某一瞬间或时间段。
한	一	4급	명사	하나의.	一个的。
한꺼번에	一下子	4급	부사	몰아서 한 번에. 또는 전부 다 동시에.	合起来一次就；或指全部都同时。

KoPanda 요금제

<table>
<tr><th colspan="2"></th><th>FREE</th><th>BASIC</th><th>PREMIUM</th></tr>
<tr><td rowspan="2">결제</td><td>기간</td><td>평생</td><td>1 회</td><td>구독형</td></tr>
<tr><td>가격</td><td>무료</td><td>1,500원</td><td>9,900원 / 1개월
55,000원 / 12개월</td></tr>
<tr><td>예상문제</td><td>예상 문제 제공</td><td colspan="3">✔</td></tr>
<tr><td rowspan="2">TOPIK 모의고사</td><td>시험 환경 제공</td><td colspan="3">✔</td></tr>
<tr><td>모의고사 (5 문제)</td><td rowspan="6">-</td><td>-</td><td rowspan="6">✔
무 제 한</td></tr>
<tr><td rowspan="3">AI 평가</td><td>STT</td><td rowspan="5">✔
한 문제</td></tr>
<tr><td>점수 평가 진행</td></tr>
<tr><td>AI 답안 분석</td></tr>
<tr><td rowspan="2">AI 피드백</td><td>AI 문장, 어휘 추천</td></tr>
<tr><td>모범 답안 제공</td></tr>
</table>

예상 시장



AI 교육 플랫폼, 자연어 처리 분야

외국인 한국어 학습자

TOPIK 말하기 능력 시험 응시자

- 교육비 절감
- 배움의 진입장벽 해소
- 누적된 데이터로 양질의 교육 제공
- TOPIK 말하기 시장의 선두주자

예상 이익 규모

사업 고도화	1차 고도화	2차 고도화	3차 고도화
에듀테크 기능	일부 문제 제공 학습자 수준 맞춤형 피드백	말하기 시험 문제은행 학습자 수준 맞춤형 피드백	말하기 시험 문제은행 학습자 수준 맞춤형 피드백 비슷한 수준 학습자의 오답률 높은 문제 추천
사업 형태	베타	B2C	B2C + B2B
홍보 방안	국내 대학 버디학습 홍보	교육 박람회 참여 중국인 커뮤니티 홍보	1차 고도화 홍보방안 재외국민 교육사업 참여 해외진출 국내기업 외국인 근로자 교육
음성 인식	일부 (개선 예정)	O (고도화 진행)	O (고도화 완료)
인공지능	일부 (개선 예정)	일부 (개선 예정)	O (고도화 완료)

비즈니스 전략

마케팅

1차

- 한국어 교육 박람회 참여
- 한국어 업체 / 기관 마케팅

2차

- 중국인 유학생 커뮤니티 홍보
- 국내 대학 버디 프로그램 서포터즈
- 지역 외국인 / 다문화 센터 무료배포

브랜드 이미지 구축 & 검증

재외국민 교육기관
국내외 대학 한국어 교육 서비스
한국 기업 외국인 근로자 교육

사업 분야 확장

어휘 / 문법 교육
토픽 글쓰기 첨삭

데이터 소개

TOPIK 등급별 어휘 & 문법 데이터

등급	어휘	품사	길잡이말
1급	가게	명사	가게에 가다
1급	가격02	명사	가격이 비싸다
1급	가구02	명사	가구를 놓다
1급	가깝다	형용사	거리가 가깝다
1급	가다01	동사	학교에 가다

- 총 10,972 줄
- 동음이의어 단어 처리
- 평가 및 피드백 용 컬럼 추출

한국어 기초 사전

번호	등급	분류	대표형
47	6급	표현	-으려도
48	6급	표현	-으리라는
49	6급	표현	를 막론하고
50	6급	표현	-어 치우다
51	6급	표현	-는다는

- 어휘의 뜻풀이 및 문형 정보, 외국어 대역어 정보
- 총 74,932 줄
- 단어별 문형 정보 처리
- 외국어 대역어를 TOPIK 데이터에 연결

데이터 소개

AI Hub

인공지능 학습을 위한 외국인 한국어 발화 음성 데이터

데이터 내용

- 질문에 대한 외국인의 답변 **음성 데이터**, 이를 상/중/하로 평가한 **라벨링 데이터**
- 2022.10 제공
- 약 4,300시간 (379.34GB)
- 간 투어 제거, 특수 문자 제거

	fileName	Question	Answer	평가
135400	TH30QA230_TH0255_20211021.wav	한국에서 미용실에 얼마나 자주 갑니까? 편하게 생각하는 미용실 이용 시간은 언제입니...	저는 미용실에 자주 가지 않아요 편하게 생각하는 미용실 이용 시간은 오전인 것 같아...	중
510304	JP40QB227_JP0317_20211116.wav	스트레스를 해소하는 방법에는 어떤 것들이 있을까요? 나만의 스트레스 해소 방법이 있...	산에 가거나 바다에 가면서 자연 속에서 시간을 보낸다거나 아니면 영화관에 가서 눈물...	상
555838	CN20QB283_CN0113_20210809.wav	한국에서 쇼핑하기 좋은 장소는 어디라고 생각합니까? 그 이유는 무엇입니까?	한국에서 쇼핑하기 좋은 장소는 백화점이라고 생각합니다 왜냐하면 백화점에서는 옷 가구...	하
795407	EX30QA286_EX0935_20211104.wav	한국에서 살면서 여러분 스스로가 변한 점이 있나요? 왜 그렇게 변했나요?	한국에서 살면서 저도 모르게 성격이 급해진 것 같습니다 인터넷이 조금만 느려지면 마...	상
657209	CN10QC248_CN0054_20210731.wav	현재 한국에서 무슨 일을 하시나요?	현재 대학원에 다니고 있어요 분명 알바도 가끔 해요	상

데이터 활용 방안

- 외국인의 음성과 전사 문장으로 STT학습
- 질문과 답을 이용해 문장 주제 적합성 평가 모델 구축
- 질문에 대한 답과 어휘/문법 사전을 통해 어휘/문법 상,중,하 평가 기준 마련

사용 기술

Data

AI Hub

韩知

문화체육관광부 국립국어원

Data Engineering

pandas

NumPy

scikit learn

Front-End

Flask

HTML CSS JS

Back-End

python

Pydub

Flask

MySQL

pyAudio

Modeling

TensorFlow

GENSIM topic modelling for humans

Hugging Face

scikit learn

OpenAI

Whisper & chatGPT

KoGPT kakaobrain

kiwipiepy

Word2Vec

Symspell

Hangul-utils

KoElectra

R Rhino-morph

KoPanda 스프린트

스프린트 1
11.07 – 11.15

원활하게 작동하는 모델
음성 입력 / STT / 발화의 상.중.하 분류

스프린트 2
11.16 – 11.30

모델 개선 여부 확인
병합 모델 정상 작동 확인
발견된 오류 확인 및 추가 개선 사항 확인

스프린트 3
12.01 – 12.13

최종 모델 정상 작동 여부
백로그의 요구사항 만족 여부 확인

KoPanda

참고논문

시장 조사

한국무역협회 - 에듀테크(Edutech) 시장 현황 및 시사점, 2020-16호

글로벌 교육시장 데이터분석 업체 HolonIQ - Global EdTech in 10 Charts, 2021

STT 모델

한국음향학회지 – 발화 속도와 한국어 분절음의 음향학적 특성, 23(2), 162-172, 이숙향, 고현주, 2004

한국정보기술학회 - 개인화된 구현을 위한 음성합성 딥러닝 모델 비교, 권세영 외 3명, 2021

한국HCI학회 - 문화 차이에 따른 음성 AI 비서의 사용자 경험과 사용자 요구사항 분석, 조수인 외 1명, 2021

한국산학기술학회논문지- 국어 특성 기반의 STT 엔진 정확도를 위한 정량적 평가방법 연구, 민소연 외 3명, 2020

언어 모델

경기대학교 대학원 - 딥러닝 기반의 한국어 문법 교정 연구, 이다빈, 2021

한국정보과학회 - 딥러닝을 이용한 한국어 띄어쓰기 및 품사태깅, 의존 구문 분석 통합 모델, 김홍집 외 3명, 2019

한국정보과학회 한국컴퓨터종합학술대회 - 한국어 형태소 분석기를 이용한 KoELECTRA 기반의 개체명 인식 기법, 조우진 외 4명, 2021

한국정보과학회 학술발표논문집 - 딥러닝 기반 한국어 맞춤법 교정연구를 위한 오류 유형 분류, 구선민 외 2명, 2021.12

국제한국언어문화학회 - 딥러닝 기반 언어모델을 이용한 한국어 학습자 쓰기 평가의 자동 점수 구간 분류 (KoBERT와 KoGPT2를 중심으로), 조희련 외 4명, 2021

팀원 소개

PM
프로젝트 총괄
STT 모델 개발
데이터 전처리

언어 모델 파트 총괄
평가, 피드백 모델
Baseline 제작

평가 모델 개발
모델 성능 개선
어휘, 문법 사전 구축
웹 프론트엔드 구축

STT 파트 총괄
STT 모델
Baseline 제작
웹 백엔드 구축

임서연

피드백 모델 개발
모델 성능 개선
주제 사전 구축
웹 프론트엔드 구축

STT 모델 개발
문장 속도 평가 지표
웹 프론트엔드 구축

끝까지 경청해 주셔서 감사합니다!

KoPanda TOPIK