



키워드를 입력하십시오

나랏말싸미

: 한글 맞춤법 학습 서비스

오종현 지민석 한지훈 정유진 최예지 오찬세 박지연 박정식

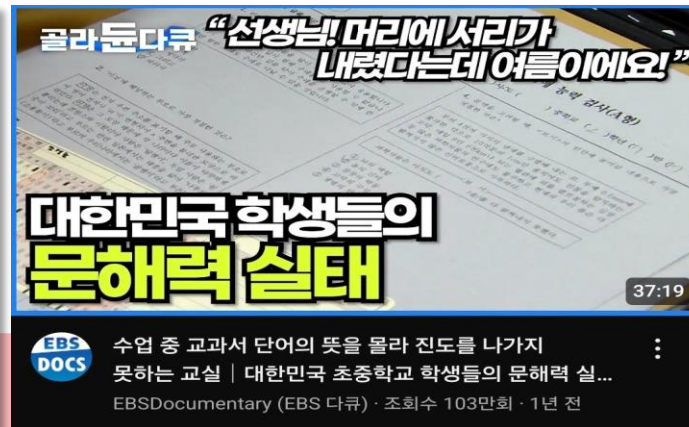
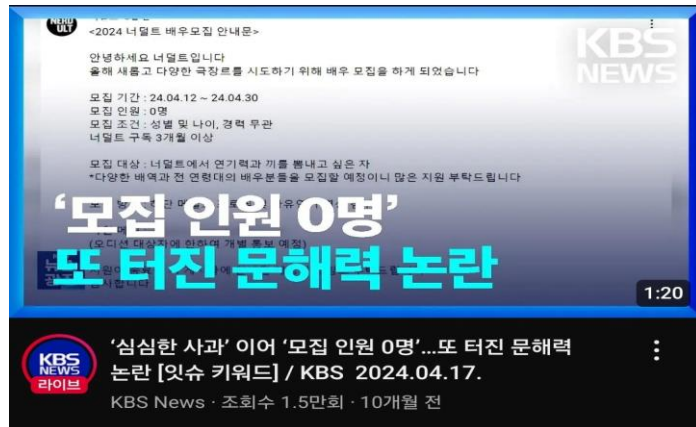
Contents

- 01 주제 선정 배경**
소목차를 작성해주세요 / 소목차를 작성해주세요
- 02 데이터 수집 및 전처리**
소목차를 작성해주세요 / 소목차를 작성해주세요
- 03 모델링**
소목차를 작성해주세요 / 소목차를 작성해주세요
- 04 서비스 아키텍처**
소목차를 작성해주세요 / 소목차를 작성해주세요
- 05 시연 및 추가 개선 방안**
소목차를 작성해주세요 / 소목차를 작성해주세요
- 06 팀원 소개**
소목차를 작성해주세요 / 소목차를 작성해주세요

1. 주제 선정 배경

주제 선정 배경

“ 점점 심각해지는 젊은 세대들의 언어 파괴 문제 ”



[로터리] 흔들리는 한국어

입력 2025-02-24 17:49:18 수정 2025.02.24 17:49:18

어 사용이 늘면서 맞춤법에 대한 관심이 줄어드는 것도 문제다. ‘되다’와 ‘돼다’를 혼
 거나 ‘낮다’와 ‘낮다’를 잘못 사용하는 사례도 많다. 이런 오류는 단순 실수가 아니라
 언어 감각이 무뎌지고 있다는 증거다. 소셜네트워크서비스(SNS)나 온라인 커뮤니티에서
 잘못된 표현이 반복적으로 노출되면서 이를 자연스럽게 받아들이는 경향도 커진다.

기존 서비스와의 차별점

기존 서비스



차별점

학습의 질 향상을 위한 **해설** 제공

사용자 맞춤 문제 **추천 시스템**

자동화된 문제 생성 방식으로
양질의 문제 생성

2. 데이터 수집 및 전처리

문법오류 데이터를 생성한 이유

기존 데이터셋



AI Hub

인터페이스(자판/음성)별
고빈도 오류 교정 데이터



국립국어원
National Institute of
Korean Language

국립국어원

언어정보 나눔터
말뭉치 데이터

해설 생성의 어려움



해설의 부재

문법오류에 대한 해설 데이터는
찾을 수 없었음

생성형 AI를 활용한
문법오류 문장 데이터와
해설 생성

데이터 생성 과정

“충분한 수의 문장과 문법오류 해설 확보를 목표로 생성”



API 선정

Gemini, open AI 등 여러 개의 생성형 AI 테스트 결과 성능이 가장 좋은 **open AI**를 활용한 프롬프트로 문장을 생성
해설은 문장에 대해 한국어에 특화된 **Clova X API**를 활용하여 생성해서 최종 데이터셋을 생성



단어쌍 수집

한국인들이 가장 많이 헛갈려하는 맞춤법 오류를 기준으로 문장을 생성

Ex) **어따** 대고 → **얼다** 대고, **바램** → **바람**



프롬프트 엔지니어링

예문 생성 시 API가 대량 데이터를 생성할 때 문장들이 **단순한 패턴**을 띄는 것을 확인

그래서 프롬프트 엔지니어링을 통해서 **조금 더 구조화된 명령**을 넘겨주어 좀 더 완성도 있는 데이터 생성

최종 데이터셋 Attributes

Content-Based Recommendation

Category

각 문법 오류의 유형

- > 0: 문법적으로 틀린 단어/구절 (몇일 / 며칠)
- > 1: 문법적으로 틀린 단어/구절은 아니지만 용례가 다른 경우 (결제 / 결재)
- > 2: 띄어쓰기에 문제가 있는 단어/구절 (온힘 / 온 힘)

Difficulty

각 문법 오류의 난이도

- > 1~5의 값으로 구성

Sequence-Based Model

Rating

문제 맞췄는지 틀렸는지 여부

- > 문제 정답 여부

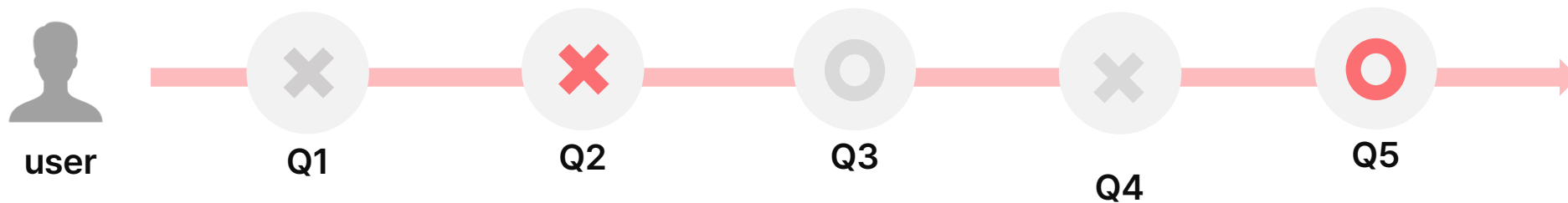
Timestamp

각 문제에 대한 사용자의 상호작용이 기록된 시간

- > Unix Timestamp

3. 모델링

모델 선정 과정



Q2. 내가 뭘 잘못했길래 그렇게 [닭달하는 vs 닥달하는] 거야?

Q5. 그녀는 아이를 [닭달하는 vs 닥달하는] 대신 혼내지 않고 타일렀다.

사용자의 **실력 향상**



최근의 유저-아이템 상호작용 영향 > **과거**의 유저-아이템 상호작용 영향



시퀀스 기반 모델 필요

모델 선정 과정

Sequence based Model

1

Bert4Rec

- Transformer 기반
- Self-attention 메커니즘
- GPU 가속 필수적, 학습 비용 높음

2

SasRec

- Transformer 기반
- Self-attention 메커니즘
- Bert4Rec 보다 가벼움

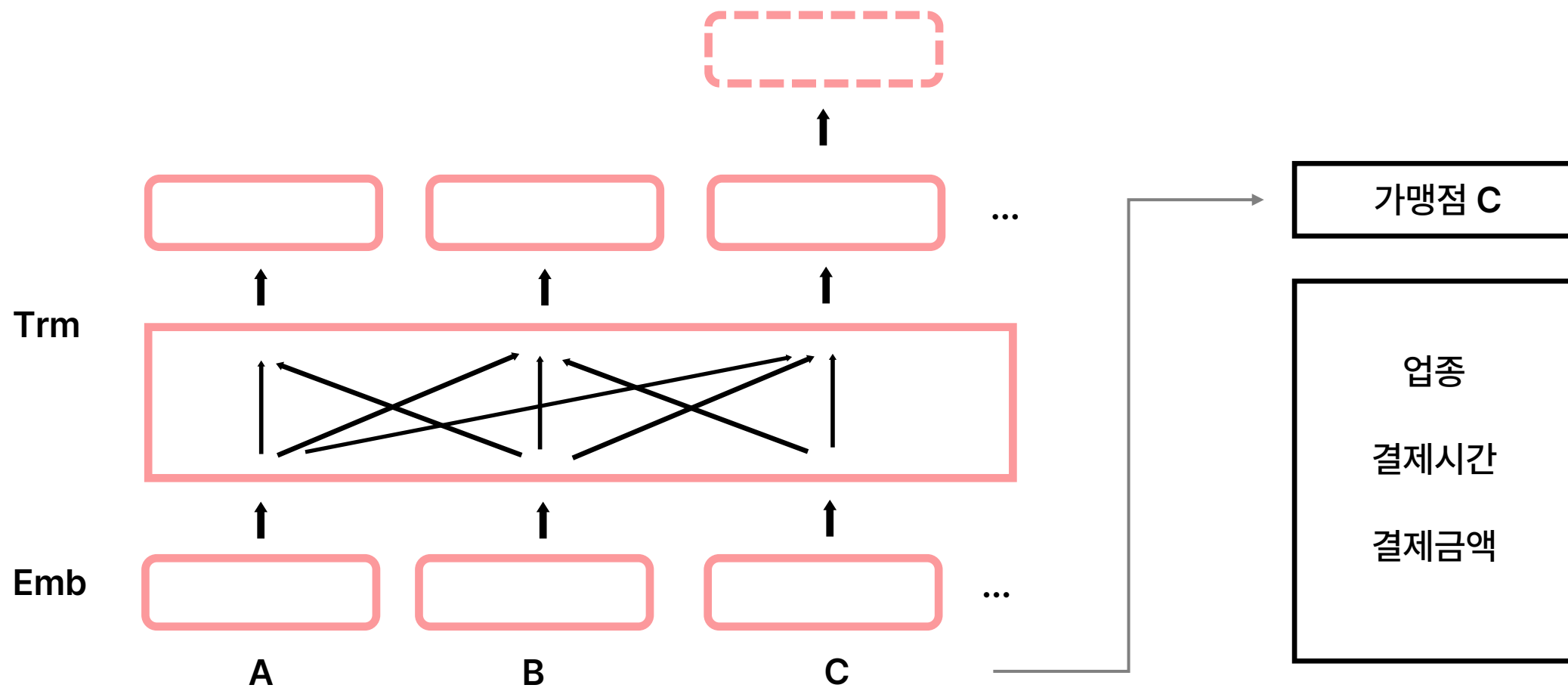
3

SasRecF

- SasRec의 확장 모델
- Recbole Library
- Features 함께 고려

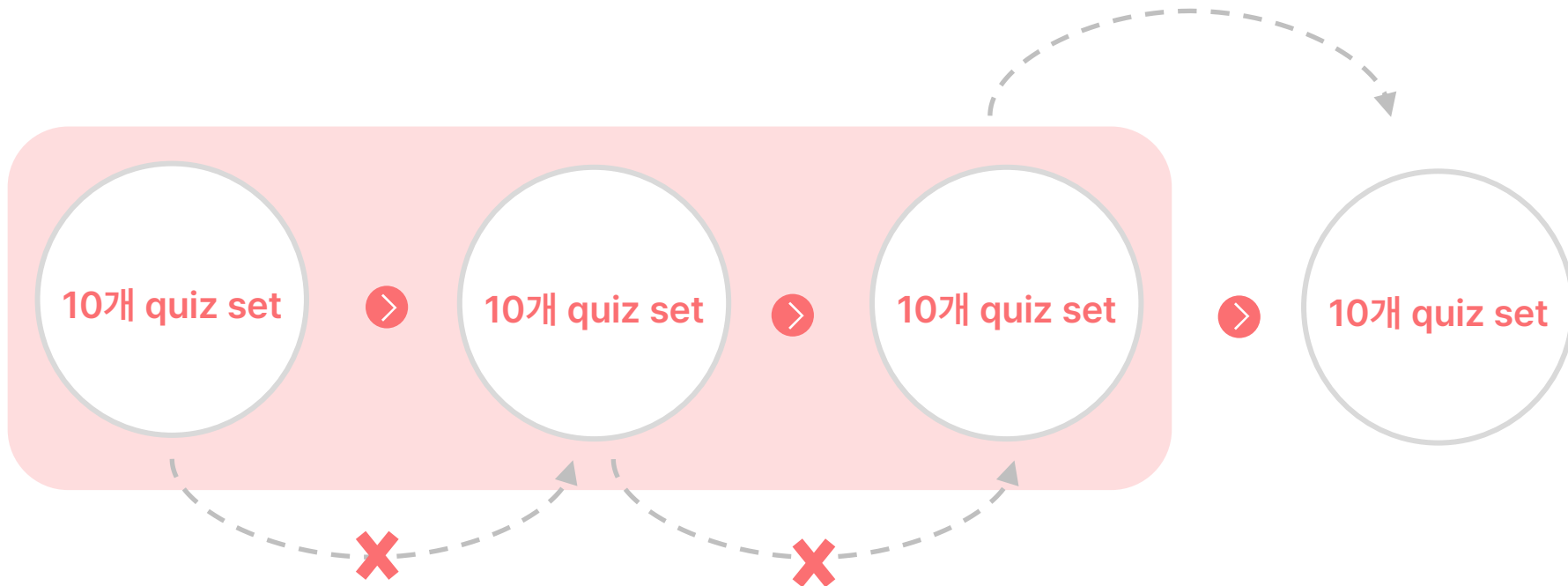
SasRecF

“SasRec에 **feature**를 반영하는 특징을 추가하여 **Recbole library**에서 만든 모델”



ColdStart 문제

30~50개의 user 행동 데이터 반영



콘텐츠 기반 추천 모델

콘텐츠 기반 추천 모델



user



Q1



Q2



Q3

	Quiz 1	Quiz 2	Quiz 3
Quiz 1	1	0.3	0.8
Quiz 2	0.3	1	0.1
Quiz 3	0.8	0.1	1

⋮

...

$$\begin{aligned}
 & -1 \times [1 \ 0.3 \ 0.8] \times a_1 \\
 & \quad + \\
 & -1 \times [0.3 \ 1 \ 0.1] \times a_2 \\
 & \quad + \\
 & 1 \times [0.8 \ 0.1 \ 1] \times a_3 \\
 & \quad || \\
 & [\ A \ B \ C \]
 \end{aligned}$$

category와 difficulty feature에 대한 코사인 유사도

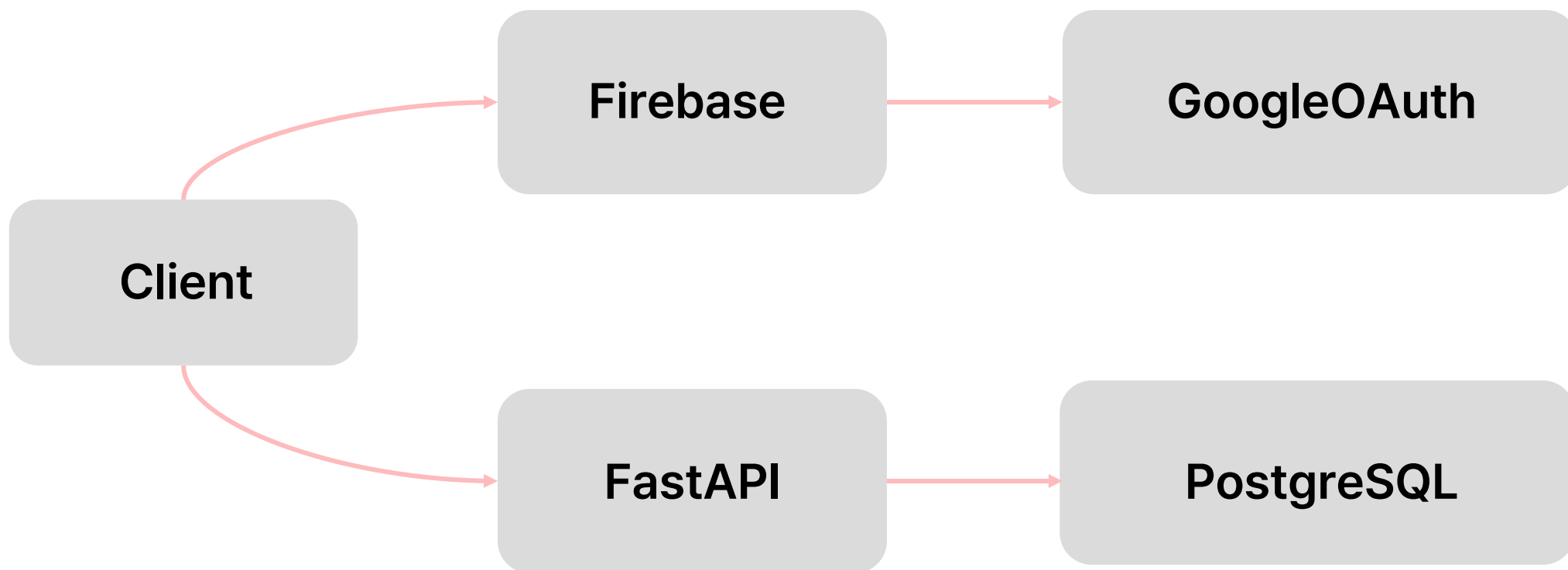
>> category 데이터 원 핫 인코딩

정답 여부, 최신성에 따른 가중치 반영

>> $a_1 < a_2 < a_3$

4. 서비스 아키텍처

서비스 아키텍처 구조



6. 팀원 소개

팀원 소개



맛쥐 종현

PM, 데이터, 모델링, 백엔드



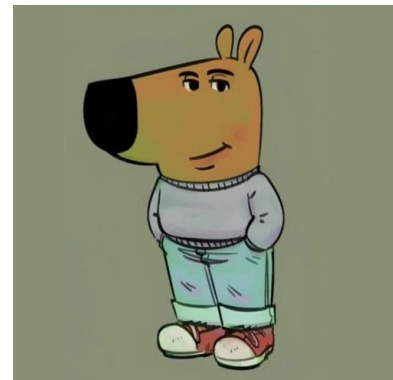
정유진

데이터, 프론트엔드



최예지

데이터 생성, 전처리



오찬세

모델링



박지연

모델링



지민석

모델링



한지훈

데이터 생성, 전처리



박정식

백엔드