

LLM 기반 한국어 문법 오류 교정 서비스

LLM-based Korean Grammar Error Correction Service

이용빈식 캡스톤 디자인 팀

이용빈, 육정훈
2024.06.12

Contents

01 팀원 소개

02 프로젝트 개요

03 제안 모델

04 시스템 아키텍처

05 데모 및 성과

06 향후 계획

*07 부록



01. 팀원 소개

팀원 소개



이용빈

팀장, 서비스 개발, MLOps 구축



육정훈

팀원, 모델 연구 및 개발, 논문 작성

02. 프로젝트 개요

프로젝트 필요성

글쓰기 - 현대인의 필수 역량

- 의사소통
- 정보 전달
- 퍼스널 브랜딩

등 글쓰기를 통해 성공적 삶을 개척 가능



프로젝트 목표

LLM 기반 글쓰기 보조 서비스 개발

서비스 주요 기능

1. 실시간 맞춤법 교정

- 문서 작성 중 실시간으로 맞춤법 오류를 감지하고 수정 제안 제공

2. 자동 글쓰기 평가

- 작성된 글에 대한 종합적인 평가 제공, 글의 질 향상 지원

3. 사용자 친화적 인터페이스

- 간편한 인터페이스로 누구나 쉽게 사용 가능

4. 다중 문서 관리

- 여러 문서의 작성, 저장, 불러오기 기능 제공

프로젝트 목표

글쓰기 보조 = GEC + AWS

- 1. GEC (Grammatical Error Correction)
 - 문법 오류 수정
- 2. AES (Automated Essay Scoring)
 - 글 수준 자동 평가


프로젝트 목표

GEC + AWS 활용한 실제 서비스 예시 (Grammarly)

Untitled document

This project aims to develop a Korean version of Grammarly, a grammar and writing assistment program for the Korean language. Currently, most grammar correction and writing assessment tools are specialized for english and do not support Korean.


3 All suggestions


 Correctness · Correct your spelling ⓘ

This ~~projaet~~ project aims to develop...

Accept


Dismiss






<

>

 Correct your spelling

assistment

 Change the capitalization

english

GEC

Performance

Text score: 80 out of 100. This score represents the quality of writing in this document. You can increase it by addressing Grammarly's suggestions.



Word count

Characters	245	Reading time	8 sec
Words	37	Speaking time	17 sec
Sentences	2		

Readability

Metrics compared to other Grammarly users

Word length	5.5	<div><div></div></div>	Above average
Sentence length	18.5	<div><div></div></div>	Above average
Readability score	42 ⓘ		

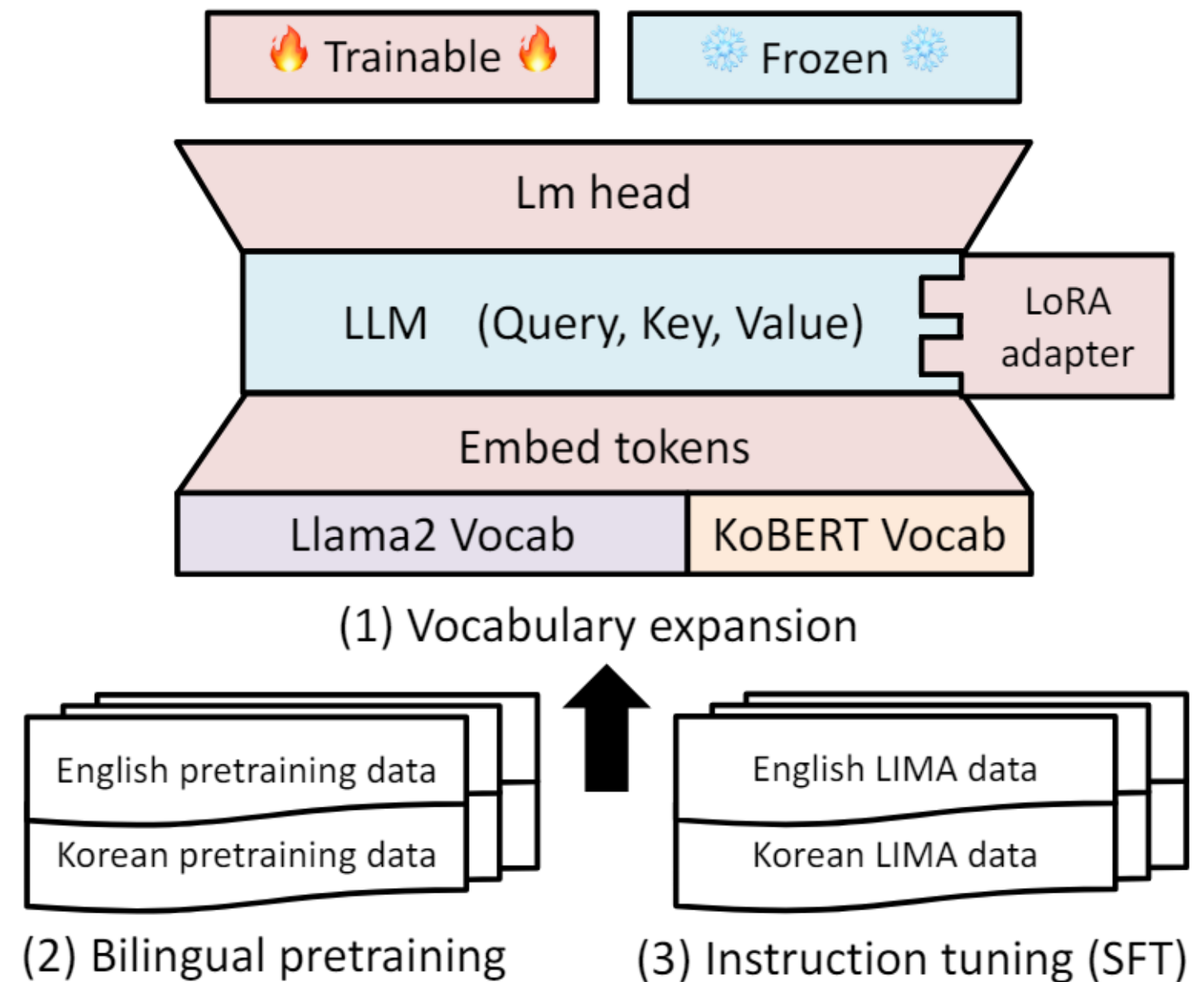
AWS

03. 제안 모델

Base Model

Blossom-13b

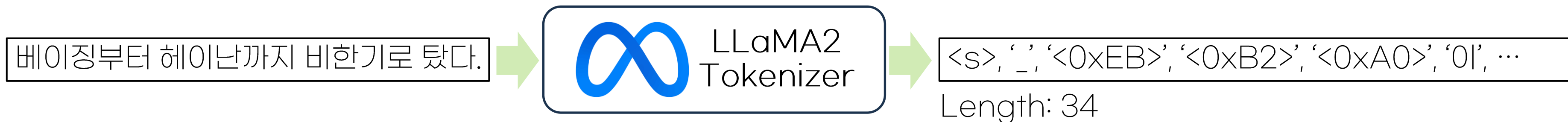
- LLaMA2-13b 기반 모델
- 한국어 Vocab 확장
 - KoBERT Vocab 이용
- 고품질의 한국어 LIMA Data SFT
 - 한국적 특성이 반영된 LIMA Data 사용



Base Model

LLaMA2-13b vs. Blossom-13b

- LLaMA2-13b
 - Pre-Train Dataset의 한글 비율: 0.06%
 - 매우 낮은 비율의 한국어 데이터
 - 한국어에 대한 모델의 낮은 이해도
 - Tokenizer Vocab의 한국어 토큰 비율: 0.0034%
 - 한국어에 대한 과도한 토큰화
 - 문법 오류가 존재할 시, 해당 단점이 더욱 부각됨



Base Model

LLaMA2-13b vs. Blossom-13b

- Blossom-13b
 - 고품질의 Korean-LIMA Dataset에서 학습
 - 고품질의 데이터를 이용한 효과적인 한국어 특성 학습
 - 기반 모델의 한국어 이해 능력 향상
 - Korean Vocab Extension
 - 한국어 토큰화 능력 향상
 - 다양한 맞춤법 오류에 대응 가능



학습 방법

SFT

- Instruction 형태로 모델에게 입력을 주는 Instruction Tuning
 - 문법 오류가 존재하는 문장, 해결된 문장의 쌍으로 구성
- Instruction의 Masking
 - Instruction의 loss 계산 제외
 - 모델의 잘못된 문법에 대한 학습 방지

Input: <s> [INST] ... “아래의 한국어 글에 대해 문법을 교정한 문장만 출력해줘. 외야하면 그때 첫사랑을 만났기 때문이다.[/INST] 왜냐하면 ... </s>
Label: <s> [INST] ... “아래의 한국어 글에 대해 문법을 교정한 문장만 출력해줘. 외야하면 그때 첫사랑을 만났기 때문이다.[/INST] 왜냐하면 ... </s>

제안하는 모델

GEC Model

- 사용 데이터셋: L1, L2
 - L1: 한국인이 작성한 맞춤법 오류 유형
 - L2: 한국어를 배우는 외국인이 작성한 오류 유형
- Input: 맞춤법 오류가 존재하는 문장
- Output: 맞춤법 오류가 해결된 문장

“<s> [INST] <<SYS>>

You are a helpful assistant.

<</SYS>>

“아래의 한국어 글에 대해 **문법을 교정**한 문장만 출력해줘.

외야하면 그때 첫사랑을 만났기 때문이다.” [/INST]

왜냐하면 그때 첫사랑을 만났기 때문이다.</s>”

제안하는 모델

AWS Model

- 사용 데이터셋: L2 글쓰기 평가
 - 외국인이 작성한 한국어 글의 Level, Score
- Input: 외국인이 작성한 한국어 글
- Output: 종합적인 수준, 점수
 - Level: 1~6급
 - Score: 0~100점

“<s> [INST] <<SYS>>

You are a helpful assistant.

<</SYS>>

“아래의 {주제}에 대한 글의 종합적인 수준과 점수를 매겨줘.

나오코씨와 빌리씨는 밥을 먹었다. 그래서 ...”

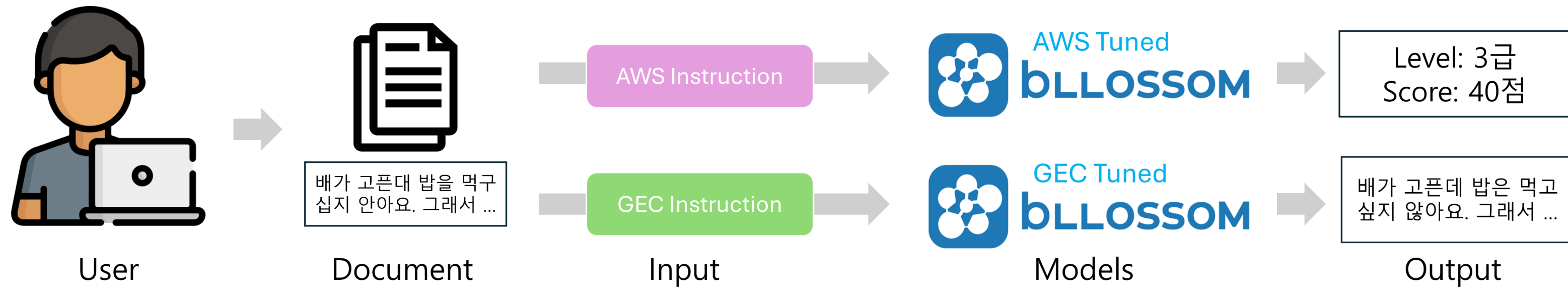
[/INST]

Level: 3급

Score: 20 </s>”

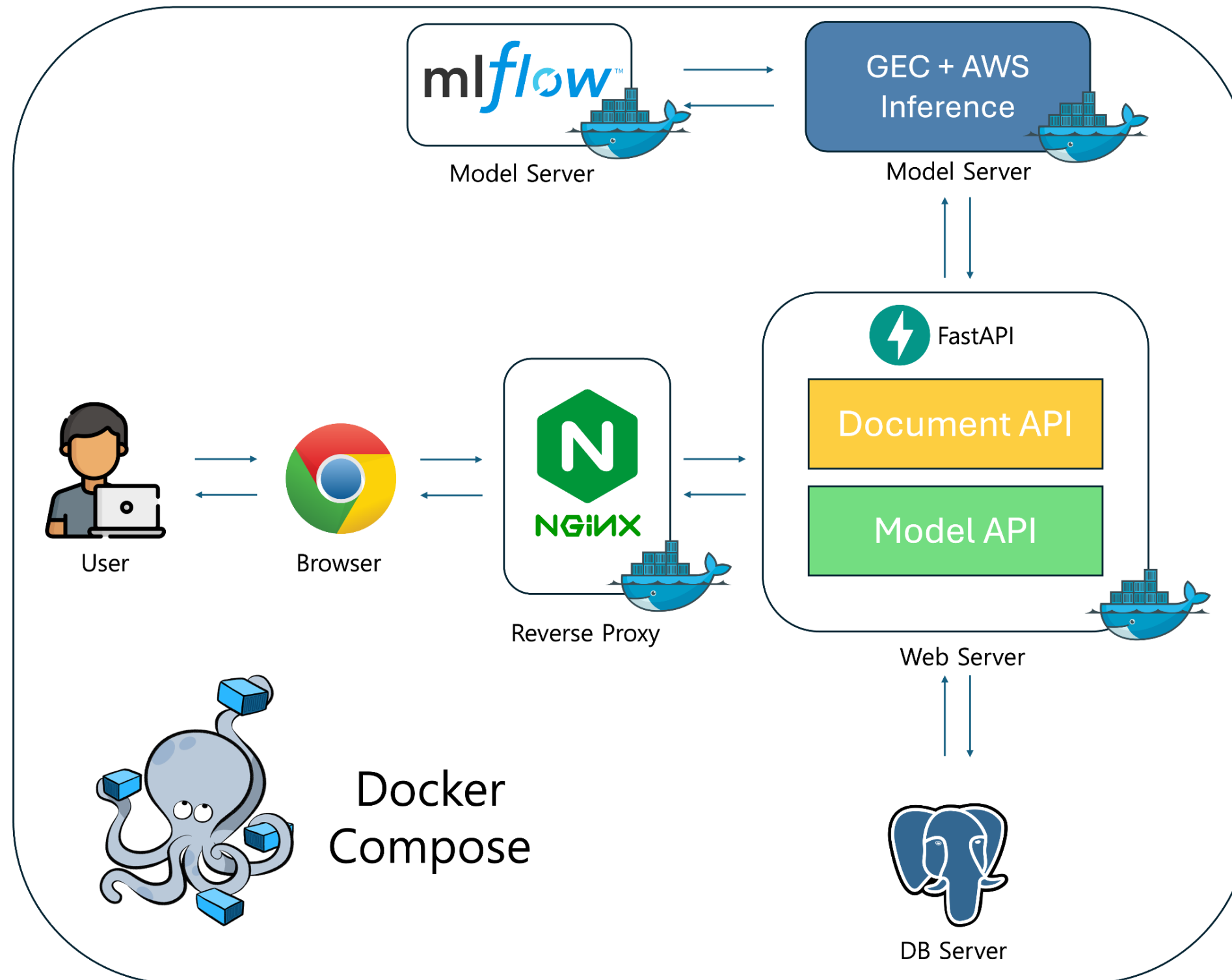
04. 시스템 아키텍처

모델 서버 개요도



<모델 서버 추론 파이프라인>

시스템 개요도



<컨테이너 기반 웹 서비스>

05. 데모 및 성과

성능 지표

GEC

- Precision
 - 존재하는 문법 오류에 대해 모델이 올바르게 고친 비율
- Recall
 - 모델이 고친 문법 오류들 중 올바르게 고친 비율
- GLEU
 - BLEU 평가 지표에서 보다 사람의 평가에 가깝게 개선한 지표

$$\begin{aligned} P &= \frac{\sum_{i=1}^n |\mathbf{e}_i \cap \mathbf{g}_i|}{\sum_{i=1}^n |\mathbf{e}_i|} \\ R &= \frac{\sum_{i=1}^n |\mathbf{e}_i \cap \mathbf{g}_i|}{\sum_{i=1}^n |\mathbf{g}_i|} \\ F_1 &= 2 \times \frac{P \times R}{P + R}, \end{aligned}$$

주요 실험 결과

GEC Model

- L1+L2
 - 두 오류 유형을 혼합한 모델의 성능이 단일 오류에 대해 학습한 모델에 비해 뛰어남

Model	GLEU	M^2		
		Pre.	Rec.	$F_{0.5}$
KoBART	33.7	44.75	14.64	31.7
BLLOSSOM	64.34	70.12	50.03	64.91

주요 실험 결과

AWS Model

- QWK(Score)
 - Quadratic Weighted Kappa
 - 예측과 정답 사이의 차이를 가중하여 반영하는 평가 지표 [-1, 1]
- ACC(Level)
 - Accuracy
 - 정확도

ACC(Level)	QWK(Score)
0.9774	0.5709

논문 성과

KCC 논문 등재

- 모두를 위한 한국어 맞춤법 교정 모델

KCC 2024 발표논문 Index

Index 확인방법

<인덱스 예시> 26A-O1-9					
26	A	O	1	-	9
6.26(수)	오전(Am)	Oral	세션번호	-	발표순서

<인덱스 예시> 28P-P8.1-13					
28	P	P	8	-	13
6.28(금)	오후(Pm)	Poster	세션번호	-	보드번호

::: 검색 :::		모두를 위한	Search	View All	검색 논문 : 1 편
논문 번호	제 목			발표자	Index
325	모두를 위한 한국어 맞춤법 교정 모델			육정훈	26A-P1.1-6

모두를 위한 한국어 맞춤법 교정 모델¹⁾

육정훈⁰¹ 신동재² 원인호² 김상민² 송승우² 김민준² 최창수² 임현석² 유한결² 송서현² 임경태²

¹한밭대학교 컴퓨터공학과, ²서울과학기술대학교 인공지능융합학과

20191780@edu.hanbat.ac.kr, {dylan1998, wih1226, sangmin6600, sswoo, mjkmmain, choics, gustjrantk,

21102372, alexalex225225, ktlim}@seoultech.ac.kr

Korean Grammar Error Correction Model for Everyone

Junghun Yuk⁰¹, Dongjae Shin², Inho Won², Sangmin Kim², Seungwoo Song², Minjun Kim², Changsu Choi², Hyeonseok Lim², Hangeol Yoo², Seohyun Song², Kyungtae Lim²

¹Dept. of Computer Engineering, Hanbat National University

²Dept. of Applied Artificial Intelligence, Seoul National University of Science and Technology

요 약

본 연구에서는 한국어 LLM을 이용한 맞춤법 교정기를 제안한다. 제안하는 모델은 공개된 한국어 어휘와 문화를 강화한 BLLOSSOM 모델을 토대로 L1, L2, L1+L2 데이터셋에서 각각 Instruction Tuning을 진행한 모델을 활용했다. 실험 결과 기존 가장 높은 성능을 보인 KoBART 모델과 비교하여 최소 9.41%, 최대 33.21% 이상 크게 향상된 성능을 보였다.

1. 서 론

본 연구에서는 공개된 생성 언어모델과 데이터를 기반으로 한국어 문법 교정 (Grammar Error Correction)을 효과적으로 진행할 수 있는 방법을 제안한다. 문법 교정은 사용자가 작성한 글에 대해 문법 오류를 찾고 이를 교정하는 작업을 의미한다. 영어의 경우 Grammarly의 사례를 토대로 문법 교정의 효율성과 사업성 모두 인정을 받아 시장이 점차 커지고 있다. 아쉽게도 한국어는 타 언어와 비교하여 문장 구조, 어순 등이 다르며, 조사와 시제의 다양성, 존칭 등 복잡한 맞춤법 특성으로 구현이 어렵고 성능이 상대적으로 낮았다. 이러한 특성 때문에 영어 등 다른 주요 언어에 비해 한국어 맞춤법 오류 교정에 대한 연구는 부족한 실정이다. 다만, 외국인에 대한 한국어 학습 수요가 증대되는 시점에 한국어 문법 자동교정 기술은 매우 중요한 기술이다[6].

현존하는 한국어 문법교정 시스템은 대부분 규칙 기반 방식의 한국어 맞춤법 오류 교정기다[7]. 이러한 시스템은 다양한 한글의 문법에 대해 규칙을 토대로 맞춤법 오류를 검출하게 된다. 하지만 한국어의 복잡한 문법과 더불어, 여간 띄어쓰기의 예외 등이 존재하므로 규칙 기반 맞춤법 검사기에는 한계가 존재한다.

이러한 규칙 기반 방식의 한계점을 해결하기 위해, 딥러닝 기반의 한글의 다양한 맞춤법 오류와 예외들에 대응할 수 있는 모델 개발이 연구되고 있다. 최근에는 한국어 GEC에 대해 훈련된 KoBART[4]가 대표적이다. KoBART 모델은 학습 데이터와 더불어 입력의 문맥을 토대로 교정된 문장을 생성하는 Encoder-Decoder 모델을 공개했다.

하지만, KoBART 모델은 상대적으로 최근 제안되고 있는 거대 Decoder 모델과 비교해 사전학습이 부족하다는 차이점이 있다.

이에 본 연구에서는 기존 KoBART에 비해 10배 이상 큰 파라미터를 가진 한국어에 특화된 최신 LLM 모델인 BLLOSSOM[1] 모델을 이용한 한국어 맞춤법 오류 교정 모델을 제안한다. 또한 기존에 존재하던 KoBART 한국어 GEC 모델과의 성능을 L1, L2, L1+L2 데이터셋에서 각각 측정하여 한국어 GEC에 최신 LLM이 얼마나 효과적으로 동작하는지 비교한다. 본 연구의 기여점을 요약하면 다음 두 가지로 소개할 수 있다.

- 한국어의 특성을 고려한 생성형 LLM을 활용해 한국어 문법 교정 모델을 제안한다.
- L1 (모국어학습자), L2 (외국인학습자)를 동시에 고려한 통합 문법교정 모델을 제안한다.

2. 관련 연구

본 연구에서 제안하는 한국어 문법교정 방법은 한국어와 한국 문화 정보를 잘 이해하는 거대모델을 토대로 한국어 문법 교정에 특화된 학습이 필요하다.


2.1 거대 한국어 언어 모델

GPT등 Transformer의 Decoder 모델 기반 LLM이 발전함에 따라 한국어에 특화된 대규모 언어 모델이 개발되고 있다. 본 연구에서는 공개된 한국어 언어모델 가운데 가장 최근 공개된 BLLOSSOM 모델을 활용했다. BLLOSSOM 모델은 LLaMA2 모델을 기반으로 하는 한국어에 특화된 LLM으로, KoBERT의 Vocab을 LLaMA 모델의 Vocab에 합쳐 LIMA 데이터로 튜닝했다. 이는 기존 모델의 한국어 토큰이 부족한 문제를 해결하는 동시에 영어에 대한 성능 하락을 최소화하는 효과적인 학습 방법이다.

1) “본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음”(2022-0-01068)

데모

표현, 구조 등 글의 세부적인 평가 지표를 제공



Chatbot

나오코씨와 빌리씨에게 사탕이 줬다.

문장단위

- 원문: 나오코씨와 빌리씨에게 사탕이 줬다.
- 수정된 문장: 나오코 씨와 빌리 씨에게 사탕을 주다.
- 수정 이유: '나오코씨와'는 올바른 한국어 이름 표기가 아니므로 '나오코 씨와'로 수정하고, '빌리씨에게'는 잘못된 조사 사용이므로 '빌리 씨에게'로 수정해야 합니다. 또한 '줬다'는 오타이므로 동사의 올바른 형태인 '주다'로 교정합니다.

글전체

표현: 글에서는 인물 간의 상호작용을 나타내는 문장을 쓰일 때, 올바른 조사와 정확한 동사 사용이 중요합니다. 예를 들어, 사탕을 주는 상황을 설명할 때 '에게' 조사를 사용하여 누구에게 무엇을 주는지 명확히 하고, 동사 '주다'를 올바르게 사용해야 합니다. 또한, 한국어에서 이름 표기는 성(姓)이나 성씨를 나타낼 때 씨를 붙여 사용하는 것이 일반적입니다.

내용: 이 문장만으로는 구체적인 상황 설명이나 인물 간의 관계가 명확하지 않습니다. 사탕을 주는 이유나 배경에 대해 좀 더 설명을 추가하여 이야기의 맥락

Type here and press enter, 질문을 입력하고 enter 버튼을 눌러주세요.

Clear history

06. 향후 계획

모델 실험 계획

1. 단일모델 GEC+AWS

- 문단 오류 GEC와 AWS 사이의 관계 실험

2. LLaMA3 기반 Blossom

- 실험의 가설이 새로운 LLaMA3 모델에서도 유효한지 실험
- 성능 향상 폭 실험

3. 근거 있는 모델의 추론

- Rule-Based 출력의 결과를 Context로 제공
- RAG 도입 등

서비스 보완 계획

1. CI/CD 파이프라인 구축

- 개발 과정을 자동화하여 소프트웨어 개발의 효율성과 품질 향상

2. 모델 추론 성능 개선

- Nvidia Triton, torchserve 등 고성능 모델 추론 서비스 도입
- 허깅페이스 기반 모델 포맷 변경

3. MLOps 구축

- 모델 모니터링 및 로깅
- 데이터 및 피드백 루프 관리

감사합니다.

하이퍼 파라미터

- Blossom-13b
 - A6000 x 1
 - Base-model: meta/LLaMA2-13b
 - Epoch: 10ep
 - Batch_size: 2
 - Learning_rate: 5e-5
 - Optimizer: AdamW
 - LoRA
 - Rank: 64
 - 학습한 layer: embedding layer, lm head layer

하이퍼 파라미터

- Blossom-70b
 - A100 x 4
 - Base-model: meta/LLaMA2-70b
 - Epoch: 5ep
 - Batch_size: 1
 - Learning_rate: 5e-5
 - Optimizer: AdamW

추가 실험 결과

- L1

Model	GLEU	M^2		
		Pre.	Rec.	$F_{0.5}$
KoBART	67.24	75.34	55.95	70.45
BLLOSSOM	82.96	90.27	77.76	87.46

- L2

Model	GLEU	M^2		
		Pre.	Rec.	$F_{0.5}$
KoBART	45.06	43.35	24.54	37.58
<u>BLLOSSOM</u>	54.47	55.21	37.11	50.3