LLM 기반 한국어 문법 오류 교정 서비스

LLM-based Korean Grammar Error Correction Service

이용빈식 캡스톤 디자인 팀

이용빈, 육정훈

2024.09.11

Contents

01. 팀원 소개

팀원 소개

이용빈

팀장, 서비스 개발, MLOps 구축

육정훈

팀원, 모델 연구 및 개발, 논문 작성

02. 프로젝트 배경

프로젝트 필요성

글쓰기 - 현대인의 필수 역량

•의사소통
•정보 전달
•퍼스널 브랜딩

등 글쓰기의 중요성 증가

글쓰기 보조 AI 기반 서비스 일반화

- Grammarly, Trinka, HyperWrite 등

grammarly

HyperWrite

프로젝트 필요성

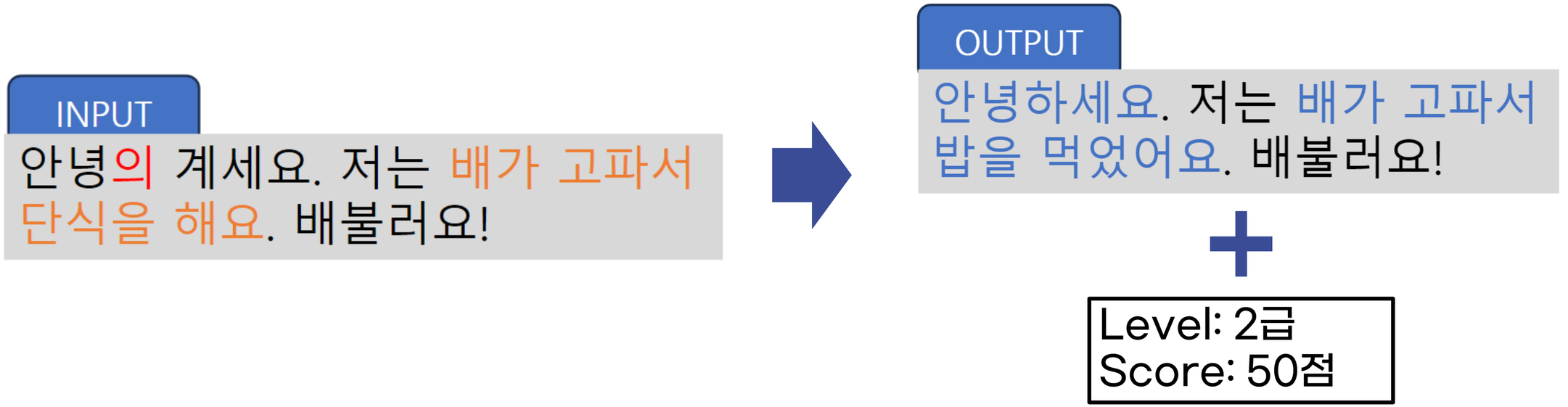
- 한국어 특화 서비스의 부재
  - Grammarly 포함 대부분의 서비스는 영문 대상
  - 일부 한국어 대상 서비스 또한 대부분이 rule-base
    - Ex) 부산대 맞춤법 검사기

INPUT

안녕의 계세요. 저는 배가 고파서 단식을 해요. 배불러요!

프로젝트 최종 목표

- 한국어 특화 LLM 기반 글쓰기 보조 서비스 개발
  - 모델이 한국어 문법을 이해하여 이에 따른 교정 제공
  - 문맥을 반영한 교정 및 평가 제공



프로젝트 최종 목표

- **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

3. 프로젝트 진행

캡스톤 1 진행 상황

- GEC 단일 수행 모델 개발

  - L1, L2 오류 유형에서 SFT된 문장 단위 GEC 모델
  - Bllossom-13b 기반 모델

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

또 나쁜 꿈은 있으지도 므르다. → GEC bLLOSSOM → 또 나쁜 꿈도 있을지도 모른다.

캡스톤 2 진행 상황

- AWE 단일 수행 모델 추가
  - L2 학습자의 문단 단위 글의 수준(Level) 및 점수(Score) 예측

ACC(Level)	QWK(Score)
0.9774	0.5709

사용 데이터셋

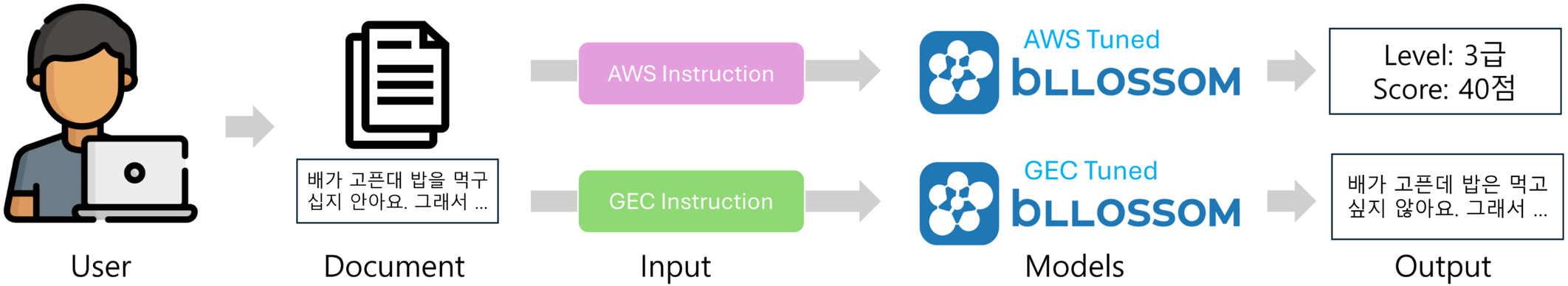
AWE(+GEC) Dataset

- 사용 데이터셋: L2 글쓰기 평가
  - 외국인이 작성한 한국어 글의 Level, Score
  - 4,011 essay, 48,937 sentences
- Input: 외국인이 작성한 한국어 글
- Output: 종합적인 수준, 점수(+교정 문단)
  - 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>"

캡스톤 2 진행 상황

- AWE 단일 모델 추가에 따른 추론 파이프라인 변경



<모델 서버 추론 파이프라인>

Multi Task로의 확장

- AWE + GEC Multi Task 모델 개발
  - 두 종류의 Input
    1. 문장 단위의 GEC
    2. 문단 단위의 AWE
  - 입력되는 두 종류의 Instruction에 따라 동작

Multi Task로의 확장

- AWE + GEC Multi Task 모델 개발

  - 단일 Task Model: AWE, GEC를 각각 수행하는 두 모델의 성능
  - Multi Task Model: AWE + GEC를 수행하는 단일 모델의 성능

모델 종류	Score (QWK)	GEC (GLEU)
단일 Task Model	0.4618	0.5076
Multi Task Model	0.4686	0.4992

* 약간의 GEC 성능 감소
* AWE 성능은 오히려 증가

서비스 측면 변화점

- 모델의 추론 속도 향상을 위한 양자화
  - 서비스를 위해, 추론 속도 증가를 위한 모델 최적화 시도
  - GPTQ 방식 사용(훈련 이후 양자화)
  - 추후 다양한 양자화 시도 예정

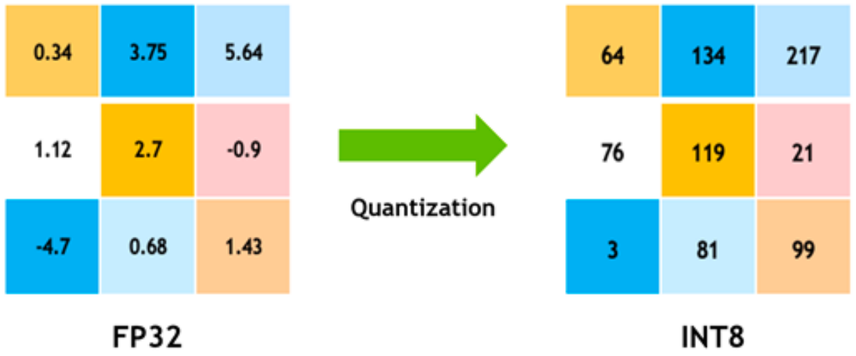


figure ref: https://medium.com/@florian_algo/model-quantization-1-basic-concepts-860547ec6aa9

서비스 측면 계획

- 원활한 서비스를 위한 추론 속도 향상
  - LLM 과 같은 인공지능 기반 서비스의 큰 문제 = 추론 속도
  - 현재 두개의 모델 추론 + 네트워크 전송 시간이 겹치면 서비스 성능 및 만족도가 낮을 것으로 예상됨
  - LLM의 추론 결과를 스트리밍 방식으로 제공하여 체감 성능 향상 예정

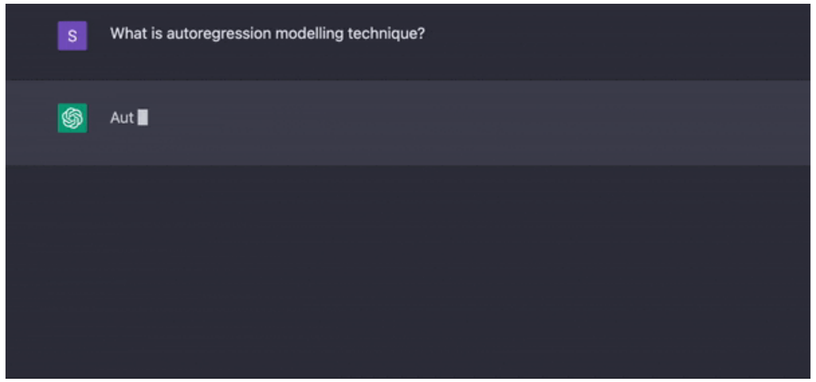


figure ref: https://blog.gopenai.com/how-to-stream-output-in-llm-based-applications-to-boost-user-experience-e9fcf582777a

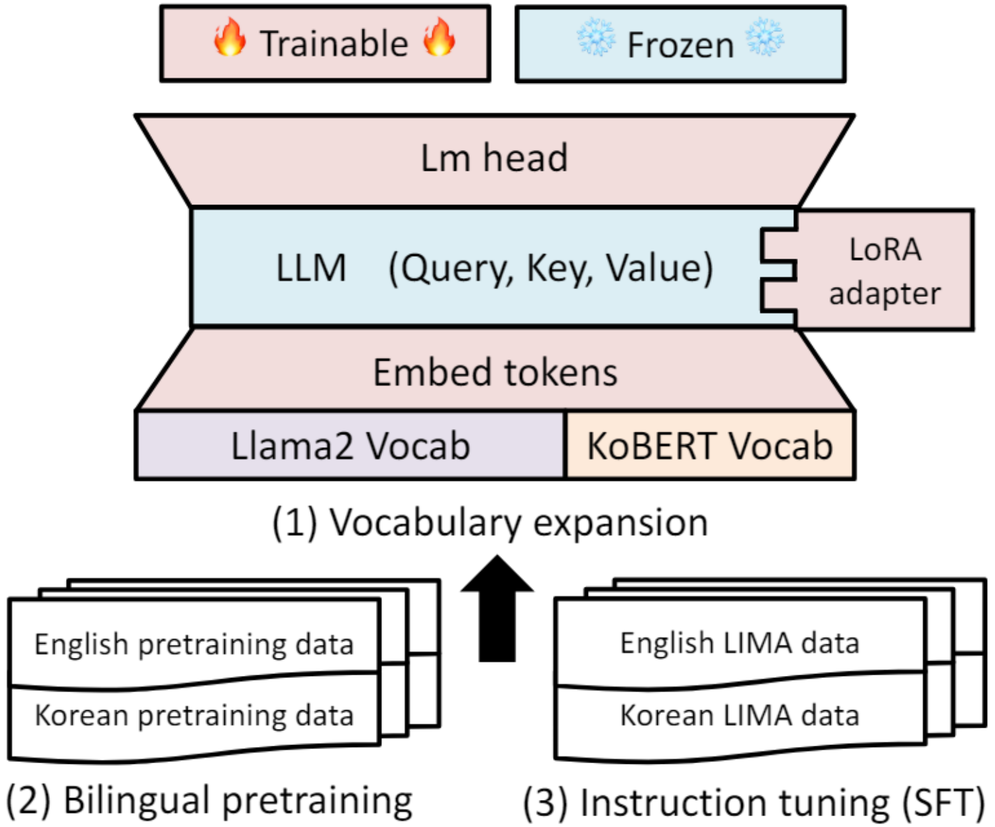
변경점 정리 및 계획

모델 실험

- LLaMA2-13b -> LLaMA3-8b
  - 최신의 open source LLM
  - 한국어 이해 성능 향상 목적

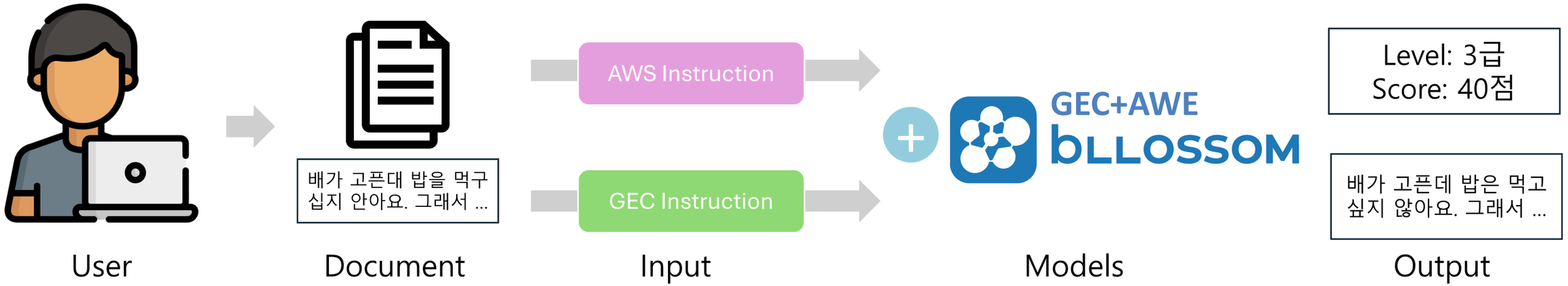
- GEC 학습 방법 실험
  - 문장 -> 문단 확장
  - Vocab 확장 유무에 따른 성능 비교
  - 다양한 학습 기법 실험
  - 근거 있는 모델의 추론(RAG 등)

- Multi-Task 모델 성능 비교
  - GEC, AWE 동시 학습 방법 비교

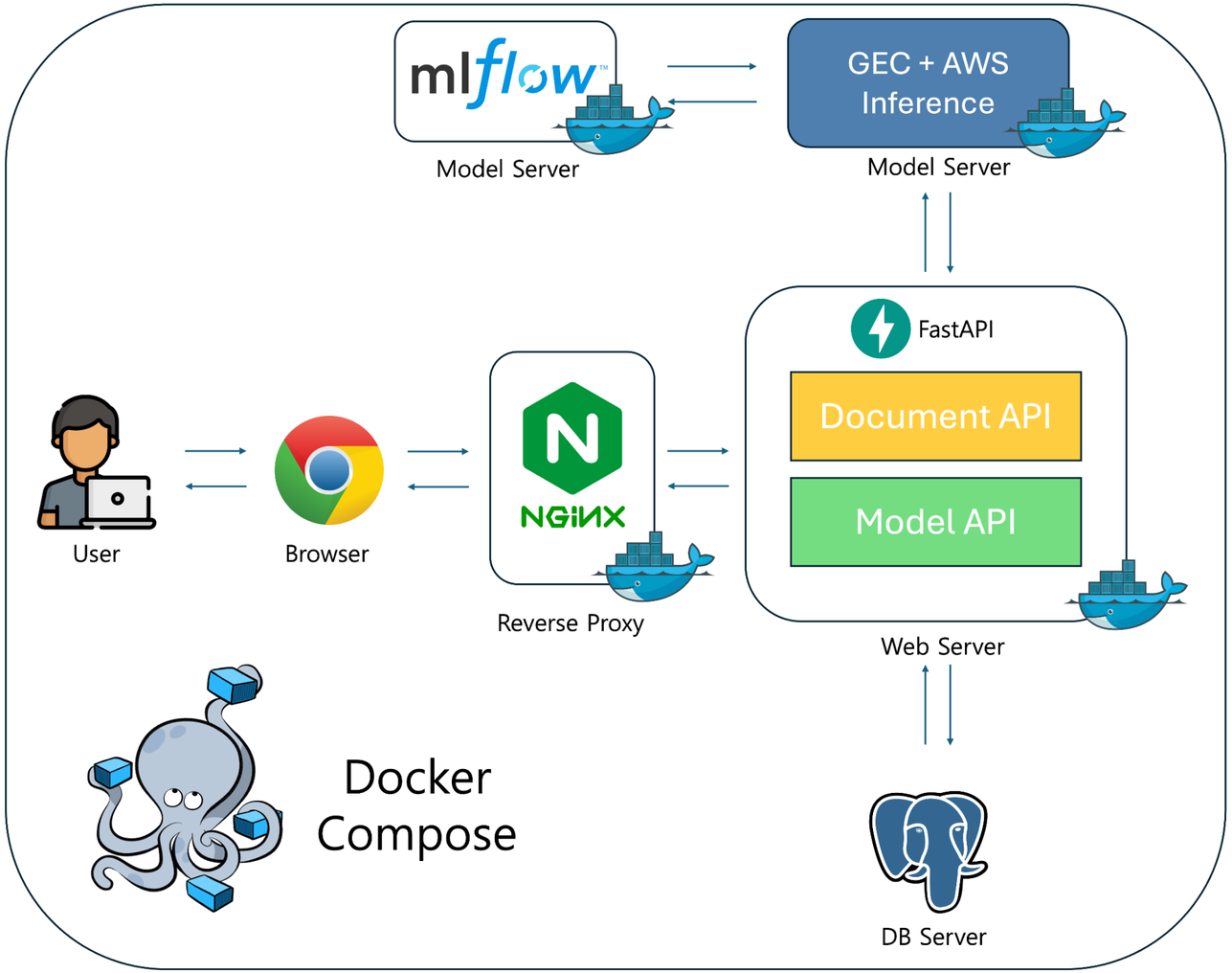


04. 시스템 아키텍처

모델 서버 개요도


배가 고픈대 밥을 먹구 십지 안아요. 그래서 ...
AWS Instruction
GEC Instruction
GEC+AWE
bLLOSSOM
Level: 3급
Score: 40점
배가 고픈데 밥은 먹고 싶지 않아요. 그래서 ...
User
Document
Input
Models
Output

시스템 개요도



<컨테이너 기반 웹 서비스>

감사합니다.

기존 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
BLLOSSOM	54.47	55.21	37.11	50.3

기존 실험 결과

- 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

사용 데이터셋

GEC Dataset

- 사용 데이터셋: L1, L2
  - L1: 한국인이 작성한 맞춤법 오류 유형
  - L2: 한국어를 배우는 외국인이 작성한 오류 유형
  - 17,487(L1) + 28,424(L2)
- Input: 맞춤법 오류가 존재하는 문장
- Output: 맞춤법 오류가 해결된 문장

"<s> [INST] <<SYS>>
You are a helpful assistant.
<</SYS>>
"아래의 한국어 글에 대해 문법을 교정한 문장만 출력해줘.
외야하면 그때 첫사랑을 만났기 때문이다." [/INST]
왜냐하면 그때 첫사랑을 만났기 때문이다.</s>"

Vocab 확장

LLaMA2-13b vs. Bllossom-13b

- Bllossom-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

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
  - BLEU의 가중치 조절

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

$$
p_n = \frac{\sum_{n\text{-}gram \in Candidate} Count_{clip}(n\text{-}gram)}{\sum_{n\text{-}gram \in Candidate} Count(n\text{-}gram)}
$$

$$
BLEU = exp(\sum_{n=1}^{N} w_n \log p_n)
$$

성능 지표

AWS Model

- QWK(Score)
  - Quadratic Weighted Kappa
  - 예측과 정답 사이의 차이를 가중하여 반영하는 평가 지표 [-1, 1]
    - 0: Random한 예측의 점수

$$\mathrm{QWK} = 1 - \frac{\sum_{i,j} W_{ij} O_{ij}}{\sum_{i,j} W_{ij} E_{ij}}$$

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2}$$

* 차이가 클 수록 가중치도 커짐

* 랜덤하게 생성된 분포와의 편차 비율

- ACC(Level)
  - Accuracy
  - 정확도

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>

IGNORE INDEX

Stage1 – GEC Dataset
39,076 Samples(Sentence)

Instruction
해당 글의 **맞춤법을 교정한 문장**을 출력해줘.

Input
외야하면 그때 첫사랑을 만난기 때문이다.

Output
왜냐하면 그때 첫사랑을 만났기 때문이다.

Stage2 – GEC + AWE Dataset
4,002 Samples(Essay)

Instruction
아래의 글의 **맞춤법을 교정**하고, 전반적인 내용의 질에 따른 **수준과 점수**를 평가해줘.

Input
하지만 빌리씨허고 나오코씨는 모두 사진기가 없엇어요. 그래서 …

Output
하지만 빌리씨하고 나오코씨는 모두 사진기가 없었어요. 그래서 …
Score: 30
Level: 4급