

한국어를 처음 배우는 외국인 교육용 paraphrasing 모델

문지원, 박민아, 송예지, 임소영, 정대균

NLP 프로젝트

Contents

01

주제 필요성

02

주제 소개

03

진행 과정

04

결과물

05

보완할 점

01 주제 필요성

"한국어 배울래요" 5개월새 외국인 14만명 몰린 회화학습앱

머니투데이 | 김유경 기자

2021.12.23 15:35

<https://news.mt.co.kr/mtview.php?no=2021122314140083596&type=1>

기사주소 복사

| 하이로컬, 출시 5개월만에 15만명 이용...국내외서 7억 규모 프리시리즈A 투자유치



자료=하이로컬

한류와 함께 커지는 한국어 교육 시장

- 한류의 열풍과 함께 한국어 교육에 대한 수요도 증가하고 있음

01 주제 필요성

한국어 교육 서비스의 부족

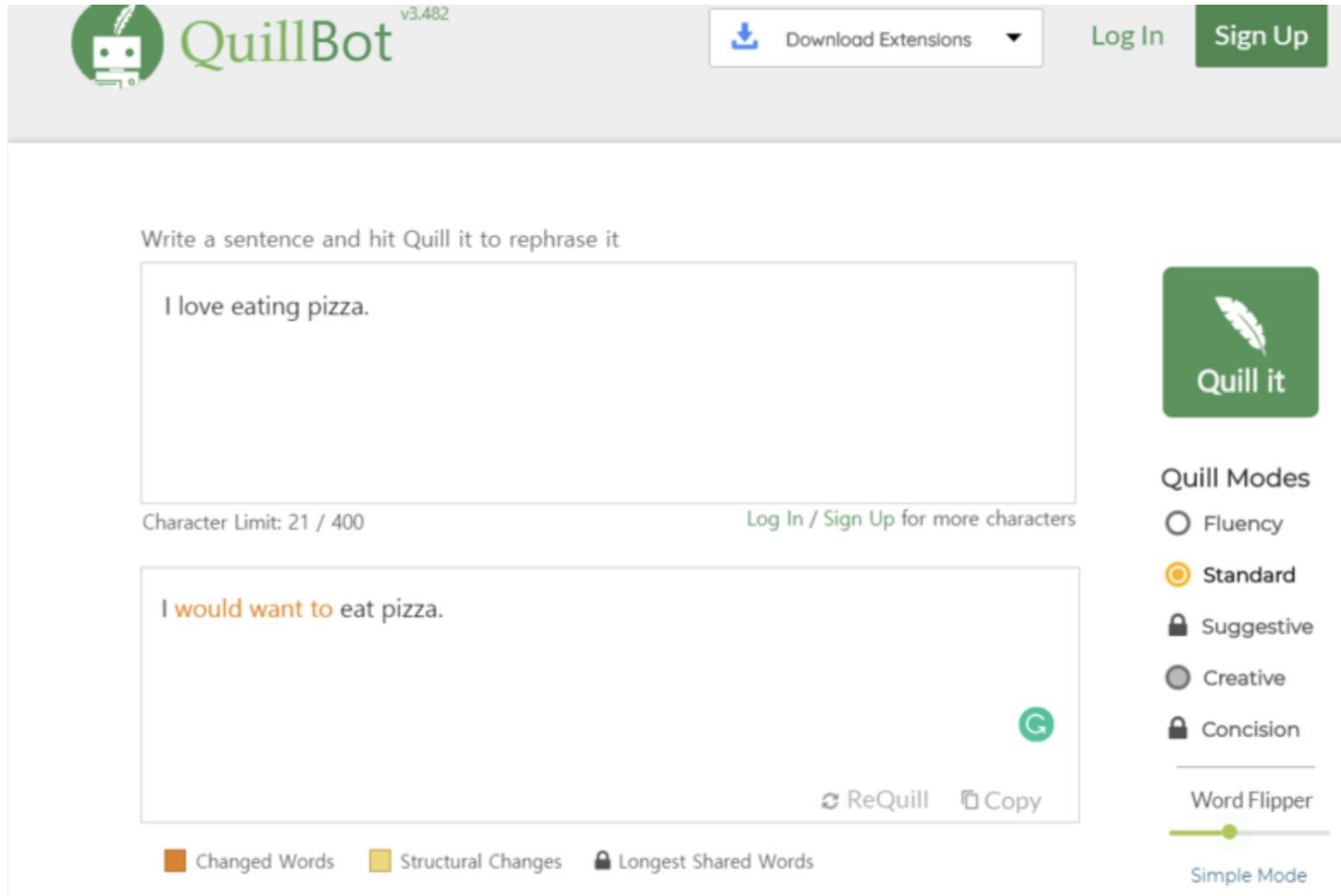
- 한국어를 새로 배우는 외국인 대상의 paraphrasing 사이트, 앱을 찾아보기 어려움
- 'Paraphrase Tool'같은 한국어 패러프레이징 사이트가 있지만, 외국인 교육용이 아님.

한국어 교육 서비스 문제점

- 사람의 검수를 거친 경우가 많음
- 한국어 : 어순에 구애받지 않는 점, 높임말의 존재
 - > 간단한 로직으로는 문장 생성이 어려움



01 주제 필요성



The screenshot displays the QuillBot v3.482 web interface. At the top, there's a navigation bar with the QuillBot logo, a 'Download Extensions' button, and 'Log In' and 'Sign Up' links. The main area prompts the user to 'Write a sentence and hit Quill it to rephrase it'. An input box contains the sentence 'I love eating pizza.' Below it, a character limit indicator shows '21 / 400' and a link to 'Log In / Sign Up for more characters'. The output box shows the rephrased sentence 'I would want to eat pizza.' with a green 'G' icon. To the right, the 'Quill it' button is visible, followed by 'Quill Modes' with options: Fluency, Standard (selected), Suggestive, Creative, and Concision. Below these are 'Word Flipper' and 'Simple Mode' options. At the bottom, there are three legend items: 'Changed Words' (orange square), 'Structural Changes' (yellow square), and 'Longest Shared Words' (lock icon).

한국어 패러프레이징을 통한 한국어 학습 서비스 개발

기존에 사람이 문장을 만드는 방식보다 더 다양하고
실생활에서 사용할 수 있는 문장들을 학습

영어 패러프레이징 사이트. 글의 수준에 따라 Fluency, Standard 등을 사용할 수 있고,
치환 가능한 단어들을 보여줌으로써 영어 공부에 활용 가능.

01 주제 필요성

문서 기반
패러프레이징
T5 모델

Fine Tuning
대화, 일상 데이터 학습



한국어 학습이
가능한 모델

01 주제 필요성

주제 의의

- 한국어를 공부하고 싶은 외국인 대상
- 한국어의 특징에 초점을 맞춰 한국어 회화 예문을 제공
 - ✓ 어순이 뜻에 큰 영향을 주지 않음
 - ✓ 상황에 따른 높임말(격식체/비격식체) 등

02 주제 소개

2023 경희대 창의자율과제 공모전 선발

최종적인 목표 (-12월 말 : AI conference 발표)

한국어 교육 서비스 앱 개발

- 한국어 구사 능력 테스트
- 수준에 맞는 대화 예문 패러프레이징 학습

2023경희대학교 『창의자율과제 공모전』

접수 기간 ▶ 2023. 4. 17. (월) ~ 2023. 5. 12. (금)

공모주제

▶ 인공지능 관련 자유 주제

참가자격

▶ 서울/국제 캠퍼스 소속 재학생 및 수료생 (학, 석, 박사과정)

지원내용

▶ 20팀 선발 후, 과제당 최대 300만 원의 재료비 지원

※ 사업단 연구비 집행 규정에 의해 연구재료비 지원

공모일정

단 계	기 간	내 용
접수 기간	4. 17. (월) ~ 5. 12. (금)	이메일 접수 (aici@khu.ac.kr)
결과 공고	5. 15. (월)	선발 결과 개별 통보 (팀장)
과제 진행	5. 15. (월) ~ 12월 말	자율적 과제 수행
멘토링	과제 진행 기간	지도 교수에 의한 멘토링 지원

제출서류 및 방법

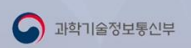


▶ 자세한 내용은 QR code 참고

유의사항

- ▶ 선정된 팀은 과제 종료 후 보고서 작성 및 제출 필수
- ▶ 12월 진행 예정인 AI Conference에서 해당 과제에 대한 발표 진행 필수
- ▶ 과제 지원비는 현금으로 지급하지 않으며, 연구재료비 집행 시 사용 가능

문의_ 031-201-5350, aici@khu.ac.kr
인공지능융합혁신인재양성사업단



03 진행 과정

01

문제정의

외국인 대상의 한국어
학습을 위한 패러프레이징
모델

02

데이터 수집 및 전처리

모두의 말뭉치, 그 외
paraphrase 데이터 수집 및
전처리

03

선행연구 조사

T5, koGPT 모델 활용
코드 및 논문 리뷰

04

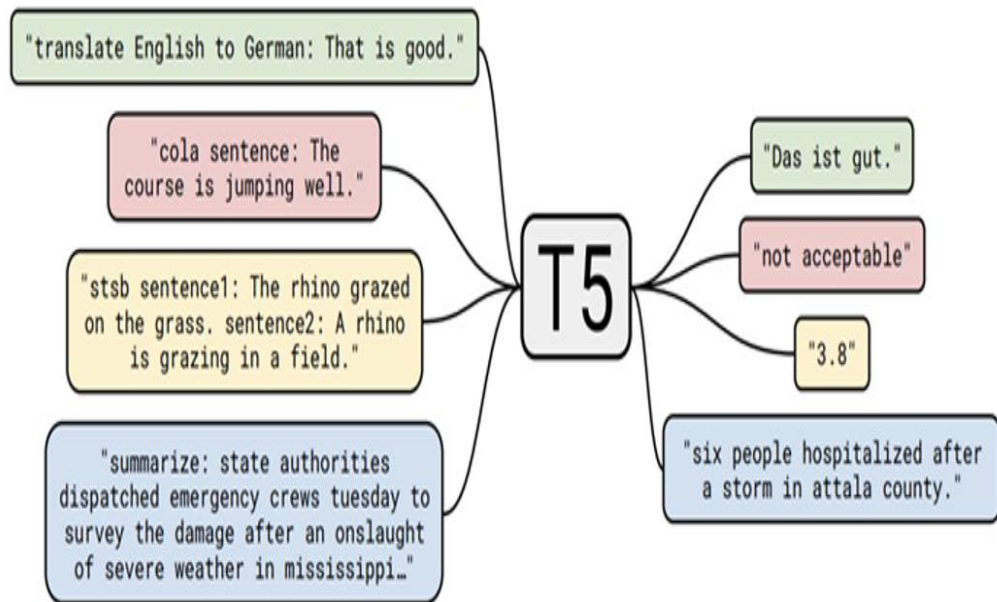
모델 학습

T5 모델
파인튜닝

03 진행 과정 - 문제 정의

Paraphrasing의 특성을 고려해 문제 정의

인코더와 디코더 구조를 모두 사용하는 seq2seq 모델 사용이 적절할 것



Paraphrase의 특성

Paraphrasing은 문장을 읽고 비슷한 문장을 생성하는 task다.



T5

Sequence to Sequence 모델인 T5가 적절하다.



KoGPT

모델의 크기 상으로는 KoGPT가 크므로 Paraphrasing에서 성능이 좋을 수 있다.

03 진행 과정 - 데이터 수집

최종 사용 데이터

- 국립국어원 유사 문장 말뭉치(모두의 말뭉치)

: (버전 1.0) 컴퓨터가 만든 유사 문장과 사람이 작성한 유사 문장으로 구성된 말뭉치

- KorSTS

: STS(semantic textual similarity)
텍스트 의미적 유사성을 평가하기 위한 데이터

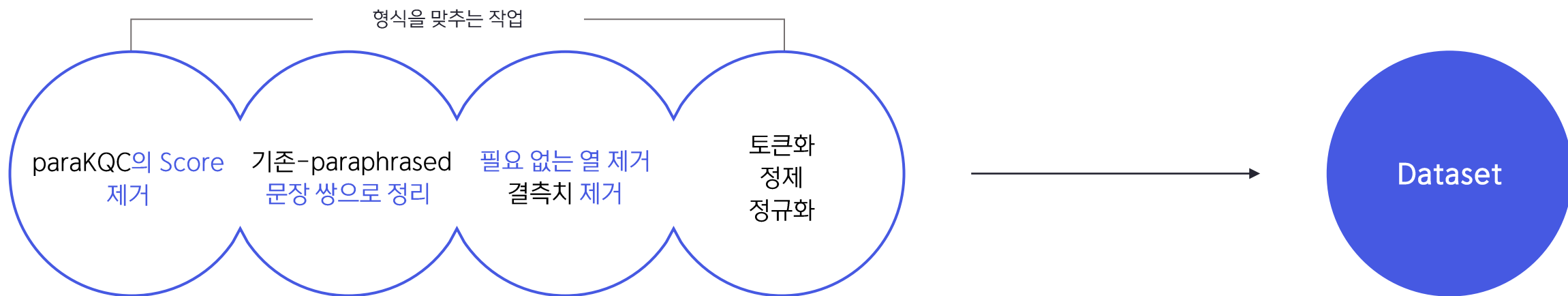
- paraKQC

:한국어 질문과 명령의 병렬 데이터 세트

“원본 문장과 유사 문장이 쌍을 이룬 형태”

	texts	pairs
0	메일 중에 안읽은 것만 지울까 다 지울까	메일을 다 비울까 아니면 안읽은 것만 지울까
1	안읽은 메일만 지워 다지워	메일을 다 비울까 아니면 안읽은 것만 지울까
2	안읽은 메일만 지워 다지워	메일 중에 안읽은 것만 지울까 다 지울까
3	다 지울까 안읽은 메일만 지울까	메일을 다 비울까 아니면 안읽은 것만 지울까
4	다 지울까 안읽은 메일만 지울까	메일 중에 안읽은 것만 지울까 다 지울까

03 진행 과정 - 데이터 전처리



03 진행 과정 - T5학습

관련 논문 조사

Paraphrasing에 활용되는 모델의 레퍼런스





- T5와 GPT 모델이 가장 많이 사용됨
- Paraphrasing은 Seq2Seq task

-> “ T5 모델을 파인 튜닝 하기로 결정 ”

패러프레이징 관련 레퍼런스 ...

↕ 2 sorts ▼

☰ 모델: T5 ▼ + Add filter

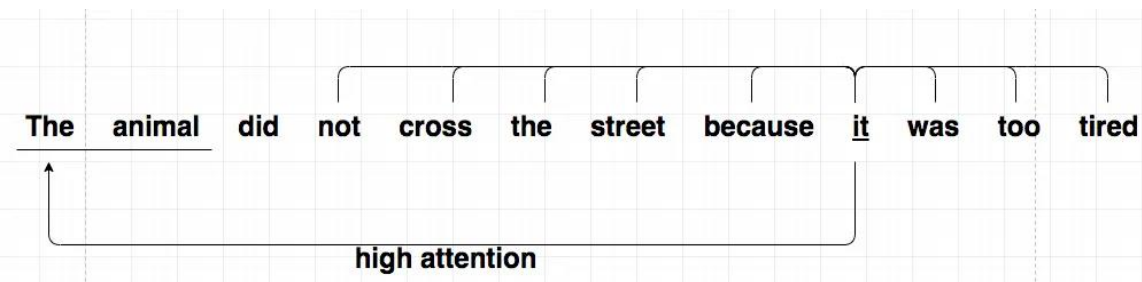
Aa 이름	Type	URL	# 우선순위	모델
 Training T5 for paraphrase generation 레퍼런스	code		0	T5
 Optimization of paraphrase generation and identification using language models in natural language processing	논문		0	T5
kolang-t5-base	code	https://github.com/seujung/kolang-t5-base		T5
A Model of Cross-Lingual Knowledge-Grounded Response Generation for Open-Domain Dialogue Systems		A Model of Cross-Lingual Knowledge-Grounded Response Generation for Open-Domain Dialogue Systems		T5
 KE-T5: Korean-English T5	code	https://github.com/airc-ke/ke-t5		T5
 Vamsi/T5_Paraphrase_Paws	code	https://huggingface.co/Vamsi/T5_Paraphrase_Paws		T5

03 진행 과정 - T5학습

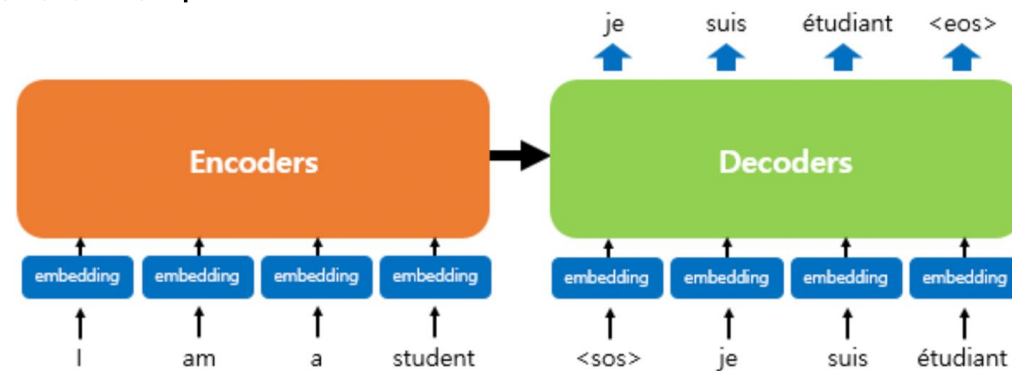
Transformer란?

- 문장 속 단어와 같은 **순차 데이터** 내의 **관계를 추적**해 맥락과 의미를 학습하는 신경망
- Attention** 구조를 촘촘하게 엮은 형태
- 바로 직전의 단어 뿐만 아니라 **문장의 다른 모든 단어**들이 문장 생성에 기여
- 인코더**는 잠재적인 표현 생성, **디코더**는 단어의 출현 확률 출력

Attention 구조: 각각의 입력 단어들이 얼마나 현재 결과에 기여하는지 나타냄

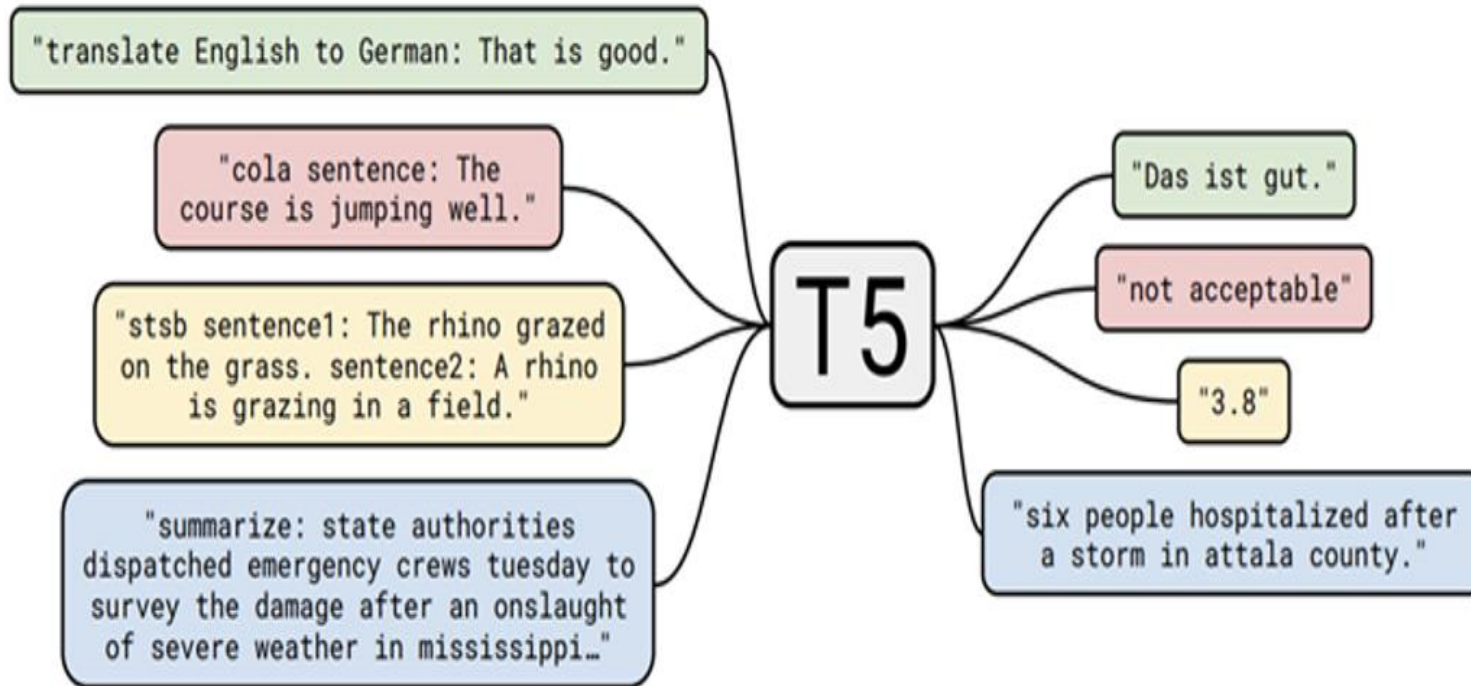


Transformer 구조



03 진행 과정 - T5학습

T5: Transformer를 기반으로 Google에서 공개

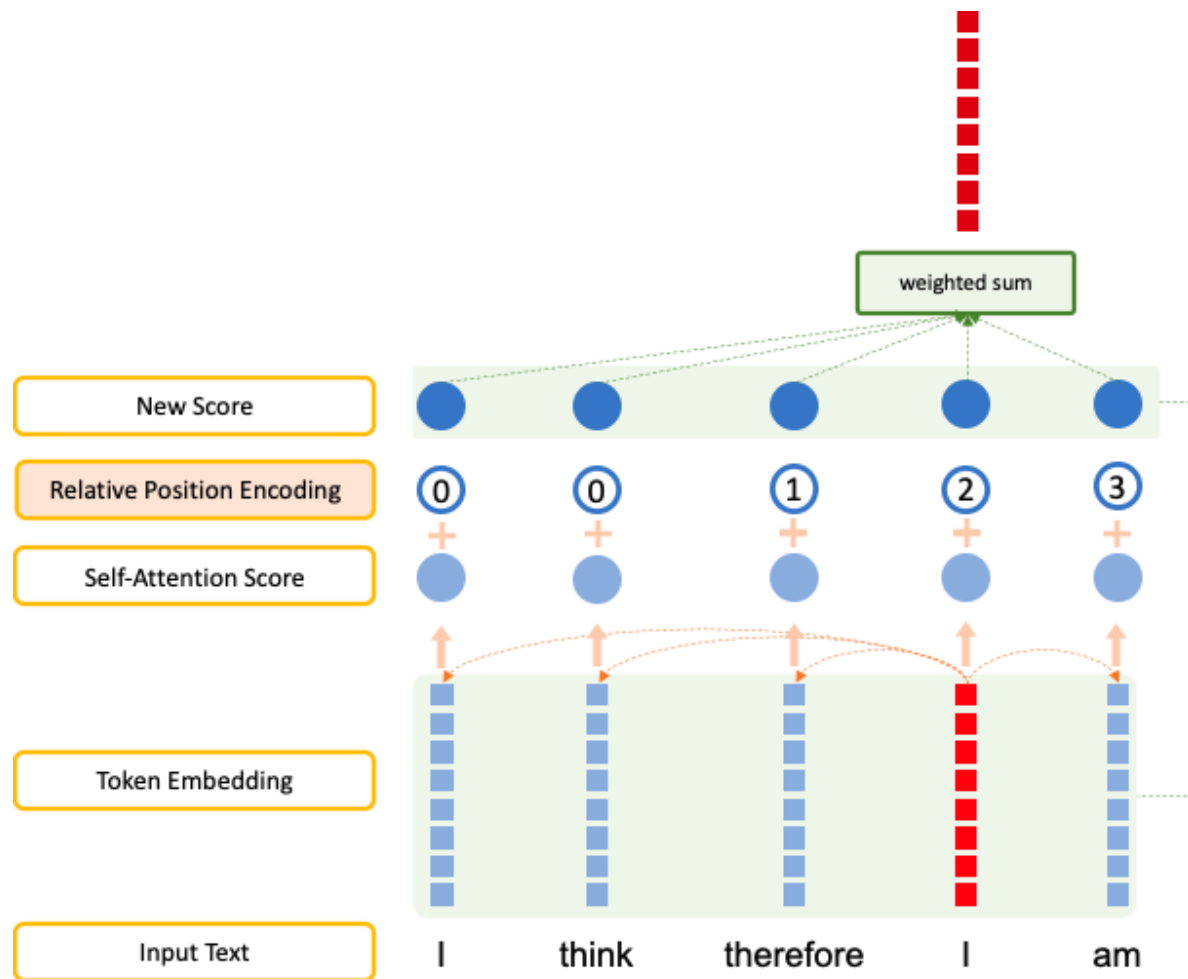


모든 task는 자연어 문장이 들어가고 자연어 문장이 나온다. => 모든 NLP task를 text-to-text로 정의

03 진행 과정 - T5학습

T5의 Relative Position Encoding

- Input으로 들어오는 각 토큰 중 특정한 범위 내의 토큰들에게 encoding 값을 부여해 가중치 계산
- 그림의 예제는 offset = 2
- 실제 T5 모델은 offset을 32에서 128로 설정해 더 많은 데이터 고려 가능



03 진행 과정 - T5모델 선정

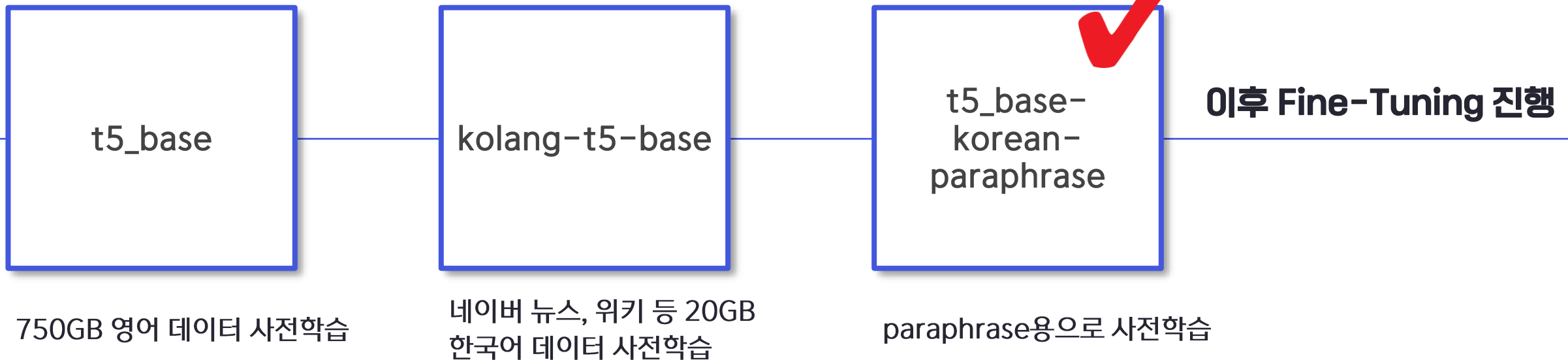
선정 모델

[Pre-trained model 선정]

lcw99 / **t5-base-korean-paraphrase**   like 2

 Text2Text Generation  PyTorch  TensorFlow  JAX  Safetensors  Transformers

 AutoTrain Compatible  License: apache-2.0



03 진행 과정 - 선정 모델의 문제점

일상대화 불가능

Original sentence:

너 이번 동아리 프로젝트 뭐할거야? 주제를 빨리 정해야하는데 생각이 잘 안 나.

Paraphrasing:



이번 동아리 활동은 어떤 것이 될 것 같은데?

이달 안에 발표될 동아리 활동 중에서 단연 돋보이는 것 하나가 너.

Original sentence:

교수님이 말하는거 이해가 잘 안 되는데 너도 그래?

Paraphrasing:



선생님이 하는 얘기 이제 나도 들을 수 있는데 어쩔 거야?

네가 교수님 말 못 들을 정도로 잘못된 게 없어 그럼 어떻게 해.

뉴스문장 가능

Original sentence:

그렇지만 화재 장비 중에 초고층 건축물 화재 발생시 초기 대응을 위한 장비는 무척이나 부족하다.

Paraphrasing:



하지만 초고층 건축물의 화재 발생시 초기에 대응하기 위한 장비는 매우 부족하다.

초고층의 건축물에 불이 났을 경우에 대비한 초기 대응 장비가 필요하지만 매우 부족하다.

04 결과물

일상 대화도 paraphrasing 가능!

[모델이 패러프레이징한 문장]

Generated Text

80	싫어하는 것이 뭐야 열대야와 높은 습도 중에서
81	네가 싫어하는 것은 더워야 아니면 추워야
82	좋아하는 계절이 여름이야 아니면 겨울이야
83	비가 오면 저기압인지 고기압인지 알려줘
84	비가 많이 오는 건 스콜인가요 장마인가요
85	태풍 진로방향은 서울이야 부산이야
86	한국과 일본 중 태풍이 어디로 향하고 있는지 궁금합니다
87	지진 후 쓰나미가 올까 아니면 화산이 폭발할까
88	이슬비와 소나기 중에 좋아하는 걸 말해봐
89	겨울에 눈이 많은 곳은 대관령이야 아니면 진부령이야

[패러프레이징 전 문장]

Actual Text

열대야와 높은 습도 중 싫어하는 것은 어떤 것인가요
더위와 추위 중 싫어하는 것을 골라주세요
좋아하는 계절은 여름과 겨울 중 무엇인가요
비가 오면 저기압일까 아니면 그 반대일까
스콜과 장마 중 비가 많이 오는 건 뭐야
태풍 진로방향이 서울일지 부산일지 알려줘
한국일지 일본일지 궁금합니다 태풍의 경로가
지진 후 쓰나미가 올지 화산이 폭발할지 알고있습니까
이슬비와 소나기 중에 뭐가 좋아
대관령과 진부령 중 겨울에 눈이 많은 곳은 어디야

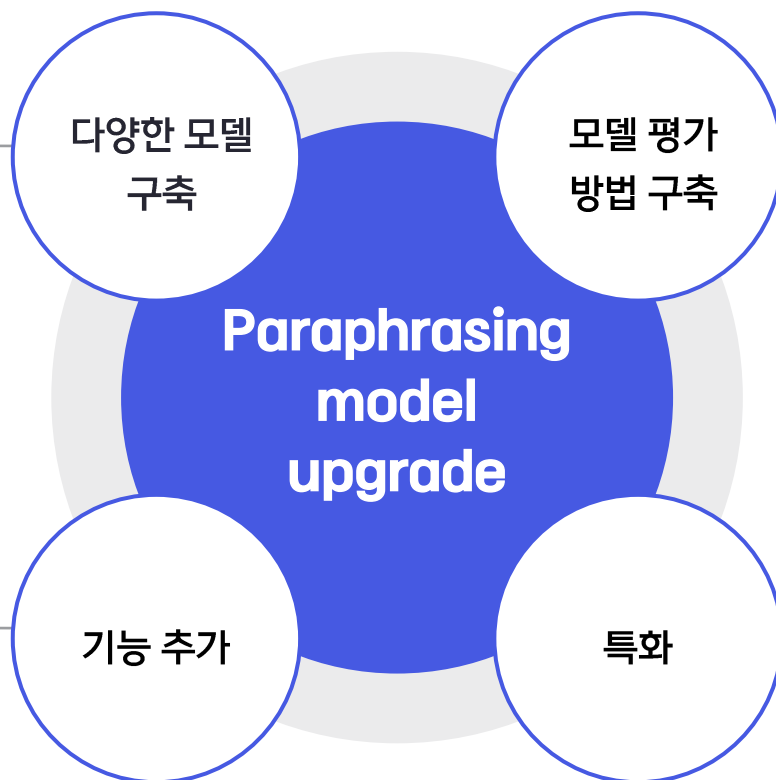
05 보완할 점

T5 모델 외 다른 모델
사용

KoGPT
Paraphrasing 모델
구축

Paraphrasing
option 추가

- 단어/구/문장 option
- 쉬운 단어/전문적인 단어 option



평가 매트릭(Rouge-L)
사용

두 문장의 LCS (Longest
Common Subsequence)를
이용

Paraphrasing model
특징 upgrade

- Paraphrasing한 부분 표시
- Paraphrasing 후 의미가 같은
이유 설명

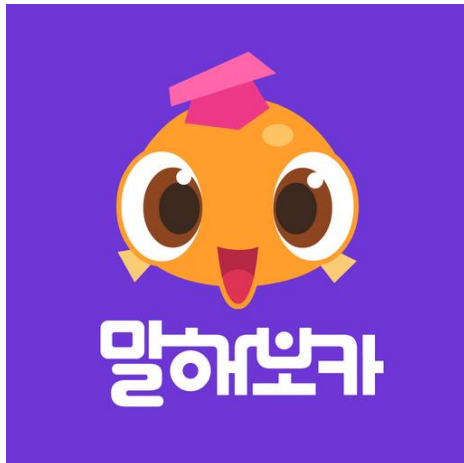
06 향후 계획

최종적인 목표 (-12월 말 : AI conference 발표)

한국어 교육 서비스 앱 개발

- 한국어 구사 능력 테스트
- 수준에 맞는 대화 예문 패러프레이징 학습

ex)



2023경희대학교 『창의자율과제 공모전』

접수 기간 ▶ 2023. 4. 17. (월) ~ 2023. 5. 12. (금)

공모주제

▶ 인공지능 관련 자유 주제

참가자격

▶ 서울/국제 캠퍼스 소속 재학생 및 수료생 (학, 석, 박사과정)

지원내용

▶ 20팀 선발 후, 과제당 **최대 300만 원**의 재료비 지원

※ 사업단 연구비 집행 규정에 의해 연구재료비 지원

공모일정

단 계	기 간	내 용
접수 기간	4. 17. (월) ~ 5. 12. (금)	이메일 접수 (aici@khu.ac.kr)
결과 공고	5. 15. (월)	선발 결과 개별 통보 (팀장)
과제 진행	5. 15. (월) - 12월 말	자율적 과제 수행
멘토링	과제 진행 기간	지도 교수에 의한 멘토링 지원

제출서류 및 방법



▶ 자세한 내용은 QR code 참고

유의사항

- ▶ 선정된 팀은 과제 종료 후 보고서 작성 및 제출 필수
- ▶ 12월 진행 예정인 AI Conference에서 해당 과제에 대한 발표 진행 필수
- ▶ 과제 지원비는 현금으로 지급하지 않으며, 연구재료비 집행 시 사용 가능



THANK YOU
