PORTFOLIO

NLP ENGINEER

오새찬 (Oh SaeChan)

PORTFOLIO

NLLB-200 Distill 350M En-Ko

NLLB-200 Distilled-350M_en2ko

The NLLB-200 model showed outstanding performance in translation task and contributed to solving problems with low-resource languages. Despite their efforts, it is still hard to run 600M or more than 1B model for those who have not enough computing environment. So I made much smaller model that expertized translaing English to Korean. you can also run it with cpu (No mixed-precision, No Quantization). [우당탕탕 개발일지]

Model

- Model: model is based on NLLB-200 600M
  - **Parameters: 350,537,728 (350M)**
  - **Encoder layers: 12 -> 3**
  - **Decoder layers: 12 -> 3**
  - FFN dimension: 4096 (same)
  - Embed dimension: 1024 (same)
  - Vocab size: 256206 (same)

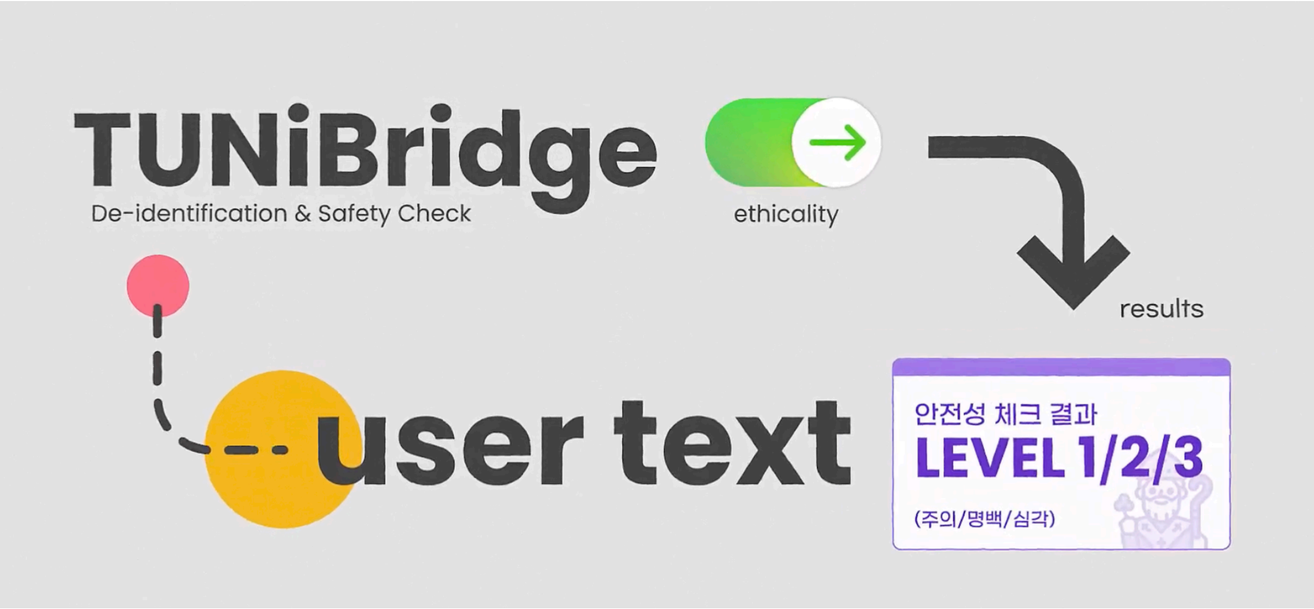
Data

- Training Data: NLLB dataset
  - created based on metadata for mined bitext released by Meta AI
  - 15M en-ko parallel dataset
  - stopes mining library
  - encoded with LASER3
- Evaluation Data: Flores-200 dataset

- Meta에서 공개한 번역모델 NLLB-200을 350M 사이즈로 경량화
- 1.3B 베이스모델을 Teacher모델로 사용하고 OnLine Distillation을 적용
- Dimension size를 유지하고 Layers의 수를 줄여서 350M 사이즈로 축소
- 가정용 GPU에서 서빙 가능하는것을 목표로 함(서빙 시 GPU용량 1.7GB를 차지, 인퍼런스 타임 2배 감소)
- Detail
  - https://github.com/newfull5/NLLB-200-Distilled-350M-en-ko
  - https://huggingface.co/dhtocks/nllb-200-distilled-350M_en-ko
  - https://devjounal.tistory.com/157

PORTFOLIO

TUNiB-Safety Model



- 텍스트 내 혐오표현 탐지 모델 개발
- 혐오표현을 종류별로 구분하고, 그에 따른 심각성을 나타냄
- 라벨러분과 지속적인 커뮤니케이션으로 데이터 공수, 데이터 관리 및 모델 트레이닝 진행
- 개인정보 필터링 모델 개발 (주민번호, 휴대폰번호, 계좌번호 등)
- Detail
  - https://www.youtube.com/watch?v=3foM20j3c_0

PORTFOLIO

TUNiB-Electra Model

TUNiB-Electra

We release several new versions of the ELECTRA model, which we name TUNiB-Electra. There are two motivations. First, all the existing pre-trained Korean encoder models are monolingual, that is, they have knowledge about Korean only. Our bilingual models are based on the balanced corpora of Korean and English. Second, we want new off-the-shelf models trained on much more texts. To this end, we collected a large amount of Korean text from various sources such as blog posts, comments, news, web novels, etc., which sum up to 100 GB in total.

You can use TUNiB-Electra with the Hugging Face transformers library.

What's New:

- Sep 19, 2021 Released a tech blog
- Sep 17, 2021 Released TUNiB-Electra.

- Electra기반 인코더 모델 개발
- 다양한 NLU 태스크에서 높은 점수를 기록함
- 각종 플랫폼에서 대용량(1TB이상) 한국어 데이터 수집 및 전처리 진행
- Detail
  - https://github.com/tunib-ai/tunib-electra
  - https://huggingface.co/tunib/electra-ko-en-base

PORTFOLIO

「하드웨어 초보자 가이드」 책 집필

RIDI

컴퓨터/IT > 컴퓨터/앱 활용

하드웨어 초보자 가이드

0명

오새찬 저

사도출판 출판

- 컴퓨터 하드웨어에대한 기초 배경지식을 담고있습니다.
- 가르치는 일에 큰 흥미가 있어서, 그 일환으로 책 집필에 착수했습니다.
- Ridi Books 페이지에서 전자책으로 출판
- Detail
  - https://ridibooks.com/books/2773000027

PORTFOLIO

AI Poet - KoGPT2

AI Poet | KoGPT2

시 생성 프로젝트의 소스코드 레포입니다.
SKT-AI에서 공개한 한국어 문장 생성 모델 KoGPT2를 활용합니다.

Data Info

- 총 1.15MB의 데이터
- 한용운 선생님 외 14명의 시인
- 시적인 가사를 사용하는 김광석 외 4명의 가수
- 신춘문예 당선작
- 교과서 빈출시 모음집

Generate Sample

- **"후회"를 키워드로 한 시 샘플**

  후회도 하련만 무엇에 걸고 살아야 할지
  빗물같은 우리의 사랑으로 나를 채운다

  후회도 하련만 무엇에 걸고 살아야 할지
  앞날은 더욱 모르노라

  맑은 날 햇살에 반짝이는 당신의 물빛,

  나의 한숨 한 방울

- SKT AI에서 공개한 KoGPT2를 SFT하여 시를 생성하는 AI 제작
- GPT2가 처음 공개되었을 당시 개발함 (2020년 6월경)
- Detail
  - https://github.com/newfull5/AI_Poet-KoGPT2

PORTFOLIO

Stereotype Detector

Korean stereotype sentence classifier

◆ Stereotype means false beliefs learned and planted by social norms or customs. For example, Programmers are good at fixing computers, programmers have poor social skills, etc...

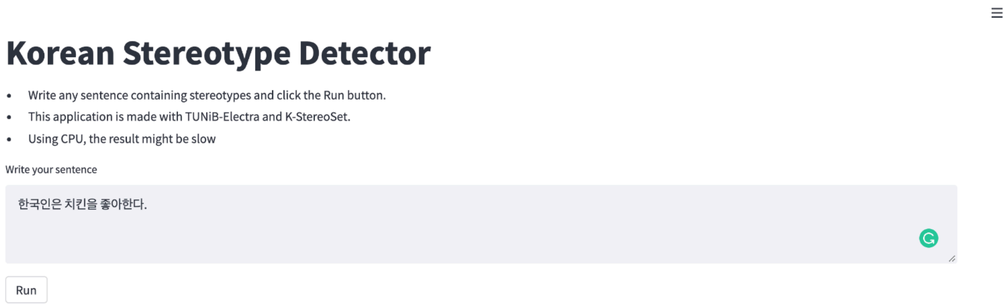
◆ This model can classify whether the text has a stereotype or not. If so, you can see what stereotypes are included (profession, race, gender, religion)

◆ You can test this models directly on the **web demo** page and also you can use this model with Hugging Face **transformers library.**

◆ This model is made using K-StereoSet with TUNiB-Electra

Web demo

- you can test this model easily in demo page
- LINK: https://share.streamlit.io/newfull5/stereotype-detector/demo.py



- 해당 텍스트가 고장관념이 포함된 문장인지 판별하는 NLU 모델 개발
- Tunib-electra와 Human label된 K-stero데이터 셋을 사용하여 SFT진행
- SteamLit을 사용하여 Web Demo페이지 제작
- Detail
  - https://github.com/newfull5/Stereotype-Detector?tab=readme-ov-file

PORTFOLIO

Solar Panel Detection

- 제주대학교 머신러닝 연구실 x 나눔에너지 프로젝트로 위성사진에서 태양광 패널을 탐지하는 RCNN 모델 개발

- Faster RCNN 모델과 위성사진에서 직접 수작업으로 라벨링한 데이터를 학습시켜서 개발

- Detail
  - https://devjounal.tistory.com/139

PORTFOLIO

ETC

- 매일 코딩 일과를 시작하기 전 알고리즘 문제(LeetCode, Programmers, 백준 등)를 하나씩 풀어 깃허브에 업로드 하는 일을 하였습니다. 현재 약 1000건 이상의 문제를 해결 했습니다. 코딩테스트에 강한 자신감이 있습니다.
  - https://github.com/newfull5/Programmers
  - https://github.com/newfull5/LeetCode
  - https://github.com/newfull5/Baekjoon-Online-Judge

- 기술 블로그를 운영하고 있습니다. 내용을 쉽게 전달하는 일의 중요함을 실감하고 꾸준히 노력하고자 하는 바람에서 시작하였습니다. 글을 꾸준히 포스팅하고자 글또5기(개발자 글쓰기 모임)에 참여하였습니다.
  - https://devjounal.tistory.com/

- 다양한 교육행사에 관심이 많습니다. Upstage AI 교육, 제주대학교 Kakao Track 등 다양한 교육 프로그램에 적극적으로 참여하였고, GDG DevFest Jeju에서 연사자로써 강연을 진행하고, 교내 교육봉사 동아리에서 학생을 대상으로 코딩교육을 진행하였습니다.