자기소개서 생성 & 한 줄 요약

쿠빅 NLP

생성2팀 프로젝트 1차 발표 [08.19]

KUBIG 13기 기다연, 이현규, 전지우

목차

0. 아이디어 회의

기다연

아이디어 (우선순위 나열)

구체적인 계획

참고자료

아이디어 (우선순위 나열)

SKT에서 개발한 KoGPT-2 (GPT-2 모델을 한글로 생성함), KoBERT, KoELEC
용?

Huggingface에서 제공하는 GPT2를 기반으로 만든 한글 GPT3 모델 사용?

1. 자기소개서 문장 생성 → 개인적으로 취업/대학입시에서 자기소개서를 작성하
   자기소개서를 앞부분/특정 키워드만 입력했을때 나머지를 다 생성할 수 있도록
2. 상담 챗봇 생성 → 거의 모든 기업체에서 상담용 챗봇 개발
3. 메일 쓰기 도와주는 생성 모델
   예) 해외의 Grammarly 서비스와 같이 메일을 쓸 때 특정 키워드/앞 문장만 ㅈ
   하는 모델
4. 카테고리별/키워드별 소설 생성

다른 옵션:

- SeqGan (생성기-판별기) 같이 사용하는 생성 모델
- GPT-3 (영어) → SKT가 올해부터 만들기 시작해서 아직 release전

이현규

1) **한국어 지역별 방언 텍스트 생성기**

   방언-표준어 전환기

- 데이터셋: AI Hub 데이터 활용

한국어 방언 발화(강원도)

https://aihub.or.kr/aidata/33979

한국어 방언 발화(경상도)

https://aihub.or.kr/aidata/33981

한국어 방언 발화(전라도)

https://aihub.or.kr/aidata/33980

한국어 방언 발화(제주도)

https://aihub.or.kr/aidata/33982

한국어 방언 발화(충청도)

https://aihub.or.kr/aidata/33984

JIT_dataset

Jejueo Interview Transcripts

https://www.kaggle.com/bryanpark/jit-dataset?select=je.dev

- 참고자료:

전지우

Text Generation(문자 생성)

1) 장르별 영화 스토리(줄거리) 생성

▶ **데이터 수집**

▶ **분석 방법(임시)**

▶ **참고 자료**

2) klue/sofo 등 평가앱을 이용해 신규 리뷰 생성

▼ **분석 방법(임시)**

1. 해당 앱의 리뷰 데이터 수집
2. 전처리 및 모델 생성 : 위에처럼 RNN 계열 모델이나 koGPT2 모델 사용(모델 간 비교를 통해 성능 평가)
3. 하이퍼파라미터 튜닝
4. 점수 및 시작 단어 입력 시 새로운 리뷰 생성

▼ **추가 개인 의견**

※ 가능하다면 분야별로 분류해서 리뷰 데이터를 모델에 적용하는 게 좋을 것 같습니다.

예) 화장품 리뷰 : 카테고리 분류 - 스킨케어, 메이크업, 옴므, 네일/향수 등

한국어 지역별 방언 텍스트 생성, 상담 챗봇, 신규 리뷰 생성 등 재미있는 아이디어가 많이 나왔습니다.

그 중, 저희가 이번 NLP 세션에서 진행하고자 하는 텍스트 생성 프로젝트는!

1. 프로젝트 소개

" 언제까지 자기소개서로 서류탈락할래! "

대학 입시부터 학회 지원, 취업까지,
자기소개서는 자신을 표현할 수 있는 가장 효과적인 방법입니다.
때문에 기업체나 대학에서 자기소개서를 더욱 더 중요한 평가기준으로 두고 있습니다.

하지만 모두가 *'합격 보장 100%'의 자기소개서* 를 쓸 수는 없기에
저희는 합격자기소개서 데이터를 기반으로 키워드나 첫 줄 입력시 자기소개서를 자동 생성하고,
문항별로 한 줄로 요약해주는 모델을 만들고자 합니다.

2. 구체적인 계획

1 합격자소서 데이터 크롤링

링커리어와 잡코리아에서
합격자기소개서 데이터 크롤링

2 데이터 EDA와 전처리

수집한 자기소개서 데이터를 이용한
EDA와 전처리 고민

3 KoGPT-2 모델 공부

SKT-AI 팀에서 발표한 KoGPT-2 모델 코드 공부
(GPT-2 모델을 한글 데이터로 사전 학습)
* Huggingface에서 제공하는 GPT-3 모델 대안

4 텍스트 요약 모델 공부

TextRank 모델/Attention 모델 코드 공부

2-1. 합격자소서 데이터 크롤링

크롤링이 가능한 합격자기소개서 제공 사이트 중 '링커리어'와 '잡코리아'에서 데이터를 크롤링 했습니다.

1. **링커리어** = 15,168건의 자기소개서 제공
2. **잡코리아** = 6,770건의 자기소개서 제공

2명의 팀원은 링커리어을 절반씩 나누어서, 1명의 팀원은 잡코리아에서 크롤링 작업을 진행했습니다.

각 자기소개서마다 합격연도(2021 상반기), 기업명, 직무명, 자소서 본문을 크롤링해 csv 파일로 저장했습니다.

2-2. 데이터 EDA/전처리

이번 주까지 데이터 크롤링 작업을 마치면,
각자 크롤링하며 데이터를 EDA하고 생각해본 데이터 전처리 작업에 관해 논의할 예정입니다.

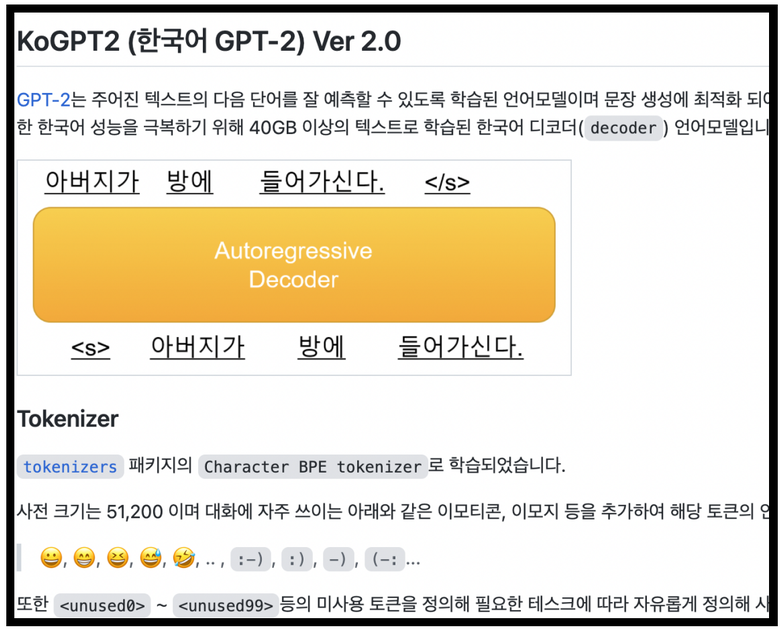
현재까지 논의해본 **데이터 전처리 방안**입니다:
- 정규표현식 이용 (한글만 남겨두고 영어/문자기호/한자 제거)
- 링커리어 합격자기소개서 데이터의 경우 문항과 답변이 함께 나오는 문제점 있음
→ 특별한 패턴이 발견되지 않아 수작업을 통해 최대한 제거
(사람마다 자기소개서 내용 내부에 문항을 표기하는 방법이 달랐기 때문에 수작업 선택)
- 띄어쓰기/맞춤법 교정 (pykospacing, pyhanspell 패키지 활용)
- 형태소 분석
- OOO 등으로 마스킹 처리된 개인정보 부분 (작성자 이름 등) 표현식으로 처리
- 회사 이름 제거
- 한글 불용어 제거

→ 최종적으로 한 줄에 하나의 문장이 들어가도록 전처리 마무리

OPTION 1

KoGPT-2

(SKT-AI 팀 배포)



KoGPT2 (한국어 GPT-2) Ver 2.0

GPT-2는 주어진 텍스트의 다음 단어를 잘 예측할 수 있도록 학습된 언어모델이며 문장 생성에 최적화 되어
한 한국어 성능을 극복하기 위해 40GB 이상의 텍스트로 학습된 한국어 디코더(`decoder`) 언어모델입니

Tokenizer

`tokenizers` 패키지의 `Character BPE tokenizer` 로 학습되었습니다.

사전 크기는 51,200 이며 대화에 자주 쓰이는 아래와 같은 이모티콘, 이모지 등을 추가하여 해당 토큰의 인

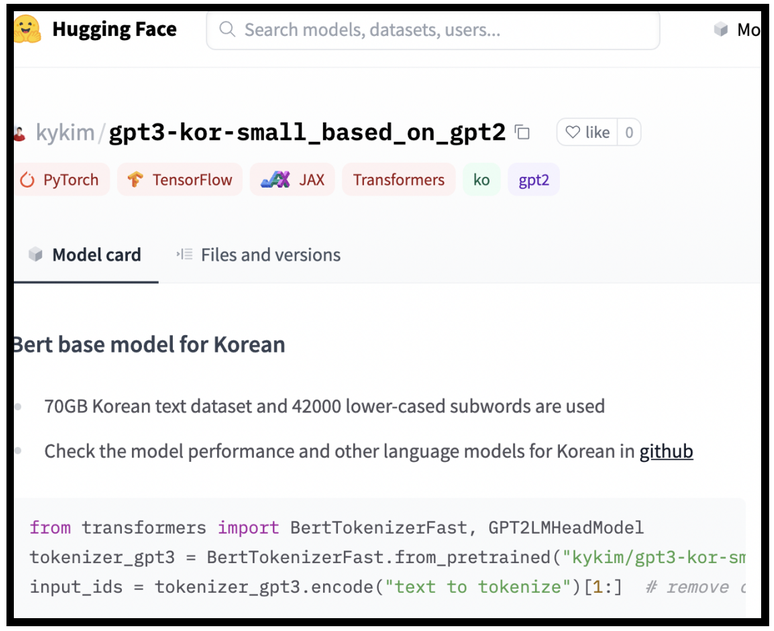
😀, 😄, 😆, 😅, 🤣, .. , `:-)` , `:)` , `-)` , `(-:` ...

또한 `<unused0>` ~ `<unused99>` 등의 미사용 토큰을 정의해 필요한 테스크에 따라 자유롭게 정의해 사

OPTION 2

GPT-2 기반 GPT-3 Small

(Huggingface)



Hugging Face Search models, datasets, users... Mo

kykim / **gpt3-kor-small_based_on_gpt2** like 0

PyTorch TensorFlow JAX Transformers ko gpt2

Model card Files and versions

Bert base model for Korean

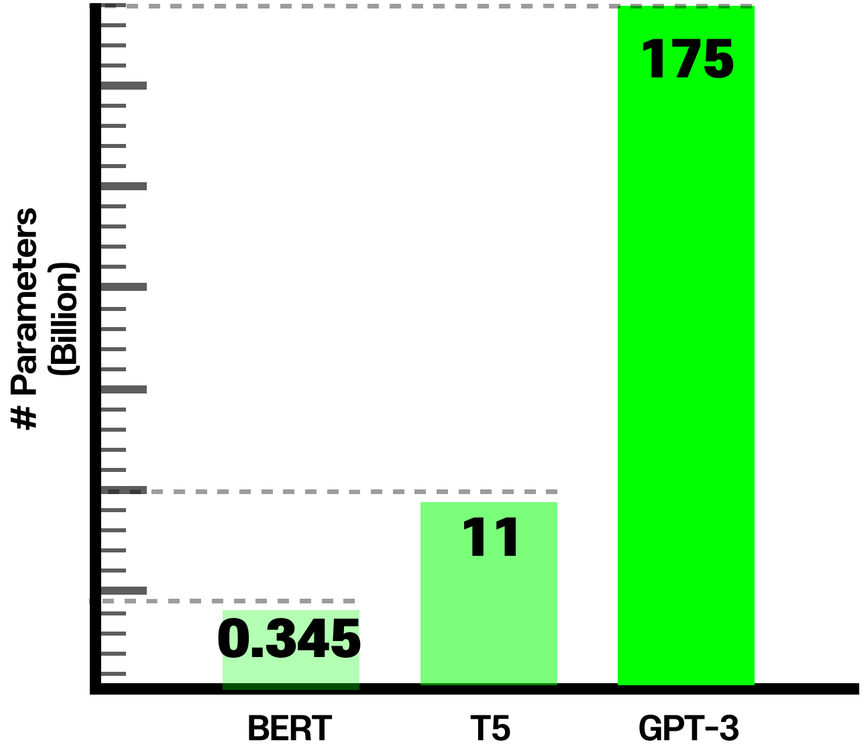
- 70GB Korean text dataset and 42000 lower-cased subwords are used
- Check the model performance and other language models for Korean in github

```
from transformers import BertTokenizerFast, GPT2LMHeadModel
tokenizer_gpt3 = BertTokenizerFast.from_pretrained("kykim/gpt3-kor-sm
input_ids = tokenizer_gpt3.encode("text to tokenize")[1:]  # remove
```

2-3. KoGPT-2 모델

GPT-2? 주어진 텍스트의 다음 단어를 잘 예측할 수 있도록 학습된 언어모델

- 문장 생성에 최적화됨
- Unsupervised 사전 학습 과정을 거침
- NLP task별로 supervised fine-tuning 과정을 거쳐야 함



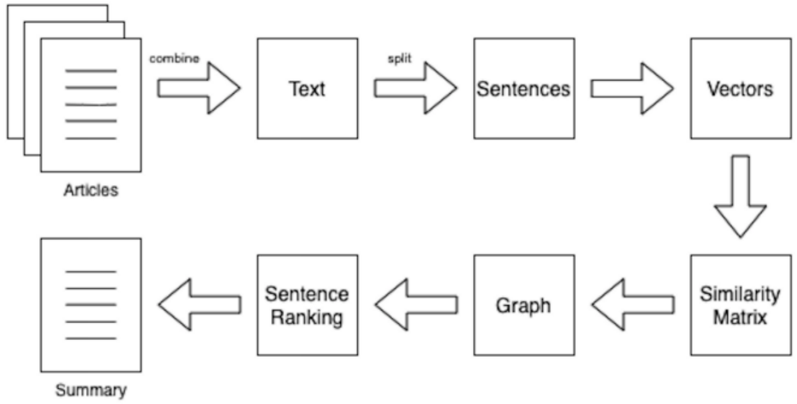
KoGPT-2?

: 기존 GPT-2 모델의 부족한 한국어 성능을 극복하기 위해 40GB 이상의 텍스트로 학습된 디코더 언어모델

크롤링, 전처리한 합격자기소개서 데이터로 fine-tuning 과정 진행

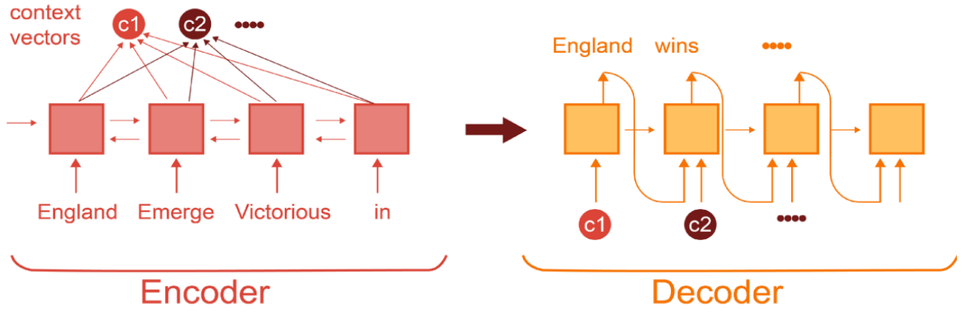
2-4. 텍스트 요약 모델

TextRank 알고리즘



: 핵심 단어를 추출하기 위해 단어간 "Co-occurence Graph (동시 출현 단어 그래프)"를 만든다. 또한 문장 단어간 유사도를 기반으로 "Sentence Similarity Graph"를 만든다.
각각 그래프에 PageRank를 학습해 각 단어마다 랭킹을 계산한다.
랭킹이 높은 키워드와 문장을 핵심 문장/키워드로 판단해 추출한다.

Attention 기반 모델



- Encoder-Decoder LSTM
- **Attention layer** 추가해서 특정

: 디코더에서 출력 단어를 예측하는 매 시점마다,
해당 시점에서 예측해야할 단어와 연관이 있는
입력 단어 부분을 좀 더 집중(attention)해서 보는 구조

3. 추후 일정

1주차

- 데이터 크롤링
- 데이터 전처리
- KoGPT-2 코드 공부

2주차

- 데이터 전처리
- KoGPT-2 모델 학습

3주차

- 하이퍼파라미터 튜닝
- 텍스트 요약 모델 학습

정기 회의

- 목요일/일요일 주2회

커뮤니케이션

- 팀 Notion 사용

이상으로 발표를 마칩니다.

질문 부탁드립니다.

감사합니다!