프로젝트 제안서

- Language & AI융합전공 202402050 최재원
- repository: github:jaeeewon/applied_nlp

데이터셋

구축

- LibriSpeech train-clean/other set 준비
- transcription을 g2p_en을 이용해 phonemes 획득
- {id, text, phonemes}의 형태로 huggingface에 데이터셋 업로드 **(완료)**
- 학습 시점에 noise를 입히고 데이터셋을 업데이트 **(예정)**
- 관련 소스코드 librispeech_phonemes.py

분석

- **dataset:** jaeeewon/librispeech_phonemes
- **split**
  - train.clean.100
    - 28,539 rows
  - train.clean.360
    - 104,014 rows
  - train.other.500
    - 148,688 rows
- 데이터 크기
  - 281,241 rows (train set only)
- 컬럼명
  - id, text, phonemes
- 예시 데이터 3개 (train.clean.100[:3])


```
{"id": "8629-261140-0000", "text": "AND GRADUALLY SUBSIDE SO THIS
CROWD YIELDED TO ITS AWE AND MAN BY MAN SANK BACK INTO HIS SEAT TILL
QUIET WAS AGAIN RESTORED AND ONLY A CIRCLE OF LISTENING FACES
CONFRONTED THE MAN WHO HAD JUST STIRRED A WHOLE ROOMFUL TO ITS
DEPTHS", "phonemes": ["AH0", "N", "D", " ", "G", "R", "AE1", "JH",
"UW0", "AH0", "L", "IY0", " ", "S", "AH0", "B", "S", "AY1", "D", " ",
"S", "OW1", " ", "DH", "IH1", "S", " ", "K", "R", "AW1", "D", " ",
"Y", "IY1", "L", "D", "AH0", "D", " ", "T", "UW1", " ", "IH1", "T",
"S", " ", "AA1", " ", "AH0", "N", "D", " ", "M", "AE1", "N", " ", "B",
"AY1", " ", "M", "AE1", "N", " ", "S", "AE1", "NG", "K", " ", "B",
```

"AE1", "K", " ", "IH0", "N", "T", "UW1", " ", "HH", "IH1", "Z", " ",
"S", "IY1", "T", " ", "T", "IH1", "L", " ", "K", "W", "AY1", "AH0",
"T", " ", "W", "AA1", "Z", " ", "AH0", "G", "EH1", "N", " ", "R",
"IH0", "S", "T", "AO1", "R", "D", " ", "AH0", "N", "D", " ", "OW1",
"N", "L", "IY0", " ", "AH0", " ", "S", "ER1", "K", "AH0", "L", " ",
"AH1", "V", " ", "L", "IH1", "S", "AH0", "N", "IH0", "NG", " ", "F",
"EY1", "S", "AH0", "Z", " ", "K", "AH0", "N", "F", "R", "AH1", "N",
"T", "AH0", "D", " ", "DH", "AH0", " ", "M", "AE1", "N", " ", "HH",
"UW1", " ", "HH", "AE1", "D", " ", "JH", "AH1", "S", "T", " ", "S",
"T", "ER1", "D", " ", "AH0", " ", "HH", "OW1", "L", " ", "R", "UW1",
"M", "F", "UH2", "L", " ", "T", "UW1", " ", "IH1", "T", "S", " ", "D",
"EH1", "P", "TH", "S"]}

{"id": "8629-261140-0001", "text": "SEEING THIS AND REALISING HIS
OPPORTUNITY FREDERICK AT ONCE ENTERED INTO THE EXPLANATIONS FOR WHICH
EACH HEART THERE PANTED THIS WILL BE OVERWHELMING NEWS TO HIM WHO HAS
CARED FOR ME SINCE INFANCY", "phonemes": ["S", "IY1", "IH0", "NG", "
", "DH", "IH1", "S", " ", "AH0", "N", "D", " ", "R", "IY1", "L",
"IH0", "S", "IH0", "NG", " ", "HH", "IH1", "Z", " ", "AA2", "P",
"ER0", "T", "UW1", "N", "AH0", "T", "IY0", " ", "F", "R", "EH1", "D",
"R", "IH0", "K", " ", "AE1", "T", " ", "W", "AH1", "N", "S", " ",
"EH1", "N", "T", "ER0", "D", " ", "IH0", "N", "T", "UW1", " ", "DH",
"AH0", " ", "EH2", "K", "S", "P", "L", "AH0", "N", "EY1", "SH", "AH0",
"N", "Z", " ", "F", "AO1", "R", " ", "W", "IH1", "CH", " ", "IY1",
"CH", " ", "HH", "AA1", "R", "T", " ", "DH", "EH1", "R", " ", "P",
"AE1", "N", "T", "IH0", "D", " ", "DH", "IH1", "S", " ", "W", "IH1",
"L", " ", "B", "IY1", " ", "OW2", "V", "ER0", "W", "EH1", "L", "M",
"IH0", "NG", " ", "N", "UW1", "Z", " ", "T", "UW1", " ", "HH", "IH1",
"M", " ", "HH", "UW1", " ", "HH", "AE1", "Z", " ", "K", "EH1", "R",
"D", " ", "F", "AO1", "R", " ", "M", "IY1", " ", "S", "IH1", "N", "S",
" ", "IH1", "N", "F", "AH0", "N", "S", "IY0"]}

{"id": "8629-261140-0002", "text": "YOU HAVE HEARD HIM CALL ME SON
WITH WHAT WORDS SHALL I OVERTHROW HIS CONFIDENCE IN THE TRUTH AND
RECTITUDE OF HIS LONG BURIED WIFE AND MAKE HIM KNOW IN HIS OLD AGE
THAT HE HAS WASTED YEARS OF PATIENCE UPON ONE WHO WAS NOT OF HIS
BLOOD", "phonemes": ["Y", "UW1", " ", "HH", "AE1", "V", " ", "HH",
"ER1", "D", " ", "HH", "IH1", "M", " ", "K", "AO1", "L", " ", "M",
"IY1", " ", "S", "AH1", "N", " ", "W", "IH1", "DH", " ", "W", "AH1",
"T", " ", "W", "ER1", "D", "Z", " ", "SH", "AE1", "L", " ", "AY1", "
", "OW1", "V", "ER0", "TH", "R", "OW2", " ", "HH", "IH1", "Z", " ",
"K", "AA1", "N", "F", "AH0", "D", "AH0", "N", "S", " ", "IH0", "N", "
", "DH", "AH0", " ", "T", "R", "UW1", "TH", " ", "AH0", "N", "D", " ",
"R", "EH1", "K", "T", "IH0", "T", "UW2", "D", " ", "AH1", "V", " ",
"HH", "IH1", "Z", " ", "L", "AO1", "NG", " ", "B", "EH1", "R", "IY0",
"D", " ", "W", "AY1", "F", " ", "AH0", "N", "D", " ", "M", "EY1", "K",
" ", "HH", "IH1", "M", " ", "N", "OW1", " ", "IH0", "N", " ", "HH",
"IH1", "Z", " ", "OW1", "L", "D", " ", "EY1", "JH", " ", "DH", "AE1",
"T", " ", "HH", "IY1", " ", "HH", "AE1", "Z", " ", "W", "EY1", "S",

```
"T", "AH0", "D", " ", "Y", "IH1", "R", "Z", " ", "AH1", "V", " ", "P",
"EY1", "SH", "AH0", "N", "S", " ", "AH0", "P", "AA1", "N", " ", "W",
"AH1", "N", " ", "HH", "UW1", " ", "W", "AA1", "Z", " ", "N", "AA1",
"T", " ", "AH1", "V", " ", "HH", "IH1", "Z", " ", "B", "L", "AH1",
"D"]}
```


- 라이선스
  - cc-by-4.0
    - LIBRISPEECH: AN ASR CORPUS BASED ON PUBLIC DOMAIN AUDIO BOOKS
      - The corpus is freely available under the very permissive **CC BY 4.0 license** and there are example scripts in the open source Kaldi ASR toolkit that demonstrate how high quality acoustic models can be trained on this data.
- 언어
  - 영어
- 목적
  - 어떤 음성에 대해 pattern matching으로 phoneme sequence를 획득했다고 하자
  - 이때, phoneme sequence를 입력으로 받아 homonym을 구분하는 영어 문장으로 바꿔주는 모델을 만들고자 함

모델 선택

- **model:** facebook/bart-large
- **paper:** BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (arxiv:1910.13461)
- 모델 정보
  - 모델 이름
    - BART (Bidirectional and Auto-Regressive Transformers)
  - 특징
    - Pretraining has two stages (1) text is **corrupted with an arbitrary noising function**, and (2) a sequence-to-sequence model is learned to **reconstruct the original text**.
    - A key advantage of this setup is the noising flexibility; arbitrary transformations can be applied to the original text, **including changing its length**.
    - BART also provides a **1.1 BLEU increase over a back-translation** system for machine translation, with only target language pretraining.
    - BART uses a **standard Tranformer-based neural machine translation architecture** which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes (see Figure 1).
    - Because BART has an autoregressive decoder, **it can be directly fine tuned for sequence generation tasks** such as abstractive question answering and

summarization.
    - 학습 데이터
        - We use the same pre-training data as Liu et al. (2019), consisting of 160Gb of news, books, stories, and web text.
            - We consider five English-language corpora of varying sizes and domains, totaling over 160GB of uncompressed text. We use the following text corpora:
                - BOOKCORPUS (Zhu et al., 2015) plus English WIKIPEDIA. This is the original data used to train BERT. (16GB).
                - CC-NEWS, which we collected from the English portion of the CommonCrawl News dataset (Nagel, 2016). The data contains 63 million English news articles crawled between September 2016 and February 2019. (76GB after filtering).
                    - We use news-please (Hamborg et al., 2017) to collect and extract CC-NEWS. CC-NEWS is similar to the REALNEWS dataset described in Zellers et al. (2019).
                - OPENWEBTEXT (Gokaslan and Cohen, 2019), an open-source recreation of the WebText corpus described in Radford et al. (2019). The text is web content extracted from URLs shared on Reddit with at least three upvotes. (38GB)
                    - The authors and their affiliated institutions are not in any way affiliated with the creation of the OpenWebText dataset.
                - STORIES, a dataset introduced in Trinh and Le (2018) containing a subset of CommonCrawl data filtered to match the story-like style of Winograd schemas. (31GB).
    - 토크나이저
        - BPE
            - Documents are tokenized with the same byte-pair encoding as GPT-2 (Radford et al., 2019).
- 모델 선택 이유
    - 우리의 목표와 유사하게 학습됨
        - **noised phonemes**(input)를 homonym도 구분하며 **sentence**(output)로 복구
            - 방대한 데이터로 학습하였기에, homonym을 구분할 수 있을 것이라고 생각함
            - **Catastrophic Forgetting**을 막고자 LoRA를 사용해볼 생각
                - 잘 되지 않는다면, noised phonemes와 sentence의 기존 데이터셋에 추가로 BART가 학습한 방식으로도 학습해볼 생각
    - 우리의 목표를 수행할 수 있을 정도의 견고한 모델 구조
    - 우리의 목표에 적절한 모델 구조
    - 우리의 목표를 쉽게 구현할 수 있는 모델 구조