1. 개요

본 문서는 2021 년 NIA-AI 학습용 데이터 구축과제(요약문 및 레포트 생성)에서 구축된 학습데이터를 활용하여 AI 모델링을 진행하기 위한 사용법을 가이드한다.
AI 모델링을 진행하기 위한 환경은 도커 이미지로 배포하므로 관련 설치는 함께 제공되는 "[요약문 및 레포트 생성 데이터] AI 모델 환경 설치가이드.pdf" 문서를 참고한다.

2. AI 모델링

2.1. AI 모델링 환경

환경 설치가이드를 통해 도커 컨테이너 실행하여 내부로 진입할 경우 아래와 같은 홈 디렉토리 구조를 확인할 수 있다.

```
root@67394bf7b5a6:/home# pwd
/home
root@67394bf7b5a6:/home# ls -al
total 48
drwxr-xr-x 1 root root 4096 Jan  3 04:01 .
drwxr-xr-x 1 root root 4096 Jan  4 05:16 ..
drwxrwxr-x 3 root root 4096 Dec 31 04:34 bert_data
-rwxr-xr-x 1 root root  268 Dec 31 02:29 docker_build.sh
-rwxr-xr-x 1 root root  248 Dec 31 06:50 docker_run.sh
drwxrwxr-x 3 root root 4096 Dec 31 04:34 json_data
drwxr-xr-x 2 root root 4096 Jan  3 04:01 logs
drwxrwxr-x 1 root root 4096 Dec 31 04:34 models
-rw-rw-r-- 1 root root 1076 Dec 31 06:52 requirements.txt
drwxrwxr-x 3 root root 4096 Dec 31 04:34 sample_data
drwxrwxr-x 1 root root 4096 Jan  3 06:15 src
drwxr-xr-x 4 root root 4096 Jan  3 06:32 temp
root@67394bf7b5a6:/home#
```

2.2. 학습데이터

{홈 디렉토리}/src 디렉토리 이동 후 sample_scripts 디렉토리의 preprocess_multi.sh 스크립트 실행(./sample_scripts/preprocess_multi.sh)을 통해 sample_date 디렉토리 내부의 원본 데이터를 이용하여 json_data, bert_data 디렉토리에 이미 생성

- ✓ sample_data : 학습 원본 데이터
- ✓ json_data : 원본 데이터(sample_data)를 전처리하여 요약 학습에 필요한 메타정보 추출, 문장 단위 분리 등의 전처리 과정을 거쳐 json 포맷으로 생성한 데이터

```
[
  {
    'src': [
      [str, str, ...],
      [str, str, ...],
      ...
    ],
    'tgt': [
      [str, str, ...],
      [str, str, ...],
      ...
    ]
  },
  {
    'src': [...],
    'tgt': [...]
  },
  ...
]
```

✓ bert_data : 모델링을 위해 토크나이징 및 PyTorch 학습을 위한 바이너리 변환한 데이터

2.3. AI 모델 학습

{홈 디렉토리}/src 디렉토리 이동 후 아래 스크립트 실행

✓ sample_scripts 디렉토리의 train_multi.sh 스크립트 실행(./sample_scripts/train_multi.sh)을 통해 학습 진행 -> GPU가 장착된 서버에서 이미 학습 완료

✓ 학습이 정상적으로 진행 완료된 경우 {홈 디렉토리}/models/MultiSumAbs_report_512 디렉토리에 바이너리(model_step_XXXXX.pt) 생성

```
root@9115bf5f42d0:/home/models/MultiSumAbs_report_512# ls -l
total 3872776
-rw-rw-r-- 1 root root 3965714631 Dec 30 09:28 model_step_15000.pt
root@9115bf5f42d0:/home/models/MultiSumAbs_report_512#
```

✓ 만일 재학습을 진행할 경우, 실행하는 호스트 서버에 GPU가 장착되지 않은 경우 train_multi.sh 스크립트의 파이썬 실행 코드의 "visual_gpus" 파라미터 값을 -1로 설정하여 학습이 가능(CPU로 학습할 경우 GPU에 비해 실행시간이 상당 오래 소요될 수 있음)

2.4. 학습된 AI 모델 평가

{홈 디렉토리}/src 디렉토리 이동 후 아래 스크립트 실행

✓ sample_scripts 디렉토리의 eval_multi.sh 스크립트 실행(./sample_scripts/eval_multi.sh)을 통해 평가 진행

✓ 평가시 실행하는 호스트 서버에 GPU가 장착되지 않은 경우 eval_multi.sh 스크립트의 파이썬 실행 코드의 "visual_gpus" 파라미터 값을 -1로 설정하여 평가 가능(CPU로 평가할 경우 GPU에 비해 실행시간이 오래 소요될 수 있음)