Summary

1. 사용한 데이터 셋

- **AI hub의 요약문 및 레포트 생성 데이터에 있는 데이터 셋을 다운로드 받아와 사용함**
- https://www.aihub.or.kr/aihubdata/data/view.do currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=582
- 데이터 구축 규모

데이터 종류	원문 규모	어노테이션 규모	결과 규모		비고
			추출요약	생성요약	
뉴스기사	27,000	59,400	14,850	29,700	2~3문장 추출
			14,850		20% 추출
보도자료	20,000	44,000	11,000	22,000	2~3문장 추출
			11,000		20% 추출
역사_문화재	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
보고서	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
회의록	34,000	74,800	18,700	37,400	2~3문장 추출
			18,700		20% 추출
사설	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
간행물	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
연설문	40,000	88,000	22,000	44,000	2~3문장 추출
			22,000		20% 추출
문학	12,000	26,400	6,600	13,200	2~3문장 추출
			6,600		20% 추출
나레이션	10,000	22,742	5,640	11,371	2~3문장 추출
			5,731		20% 추출
총계	183,000	403,342	201,671	201,671	

※생성요약은 1문장 요약으로 100글자 이하로 구축

- 학습데이터 정보

데이터 종류	학습용(Training)	검증용(Validation)	테스트용(Test)	합계
뉴스	21,600	2,700	2,700	27,000
보도자료	16,000	2,000	2,002	20,002
역사기록물 & 문화재	8,000	1,000	1,002	10,002
보고서	8,000	1,000	1,000	10,000
회의록	27,200	3,400	3,400	34,000
사설	8,000	1,000	1,000	10,000
간행물	8,000	1,000	1,000	10,000
연설물	32,000	4,000	4,000	40,000
문학	9,600	1,200	1,200	12,000
나래이션	8,371	1,000	1,000	10,371
합계	146,771	18,300	18,304	183,375

2. 데이터 량

- 균등 데이터 셋으로 구성
- AI hub에서 총 10개의 범주로 나누었기 때문에, 10개의 문서 종류 분류 범주로 나눔
- train data set : 각 분야별 문서 4000개 * 범주 10개 (총 40000개)
- valid data set : 각 분야별 문서 500개 * 범주 10개 (총 5000개)

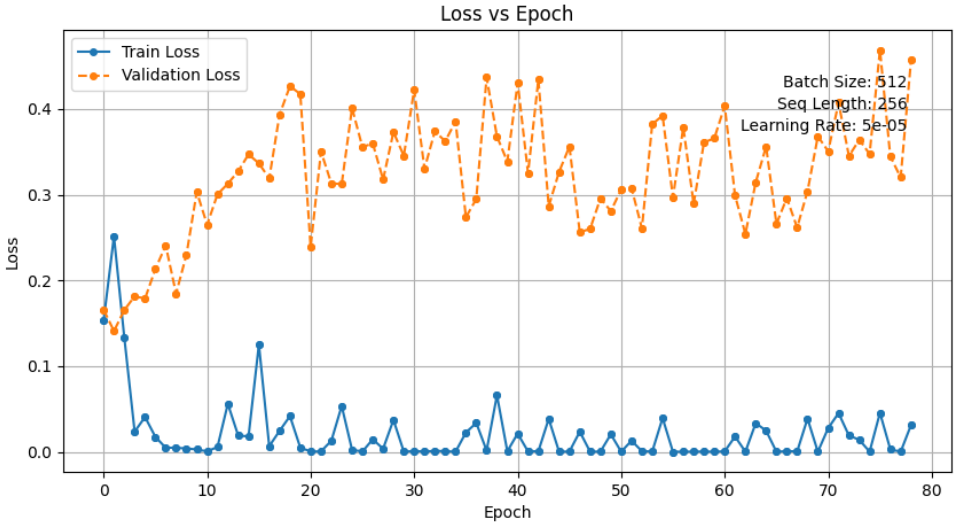
3. 학습 환경 설정 (args 설정)

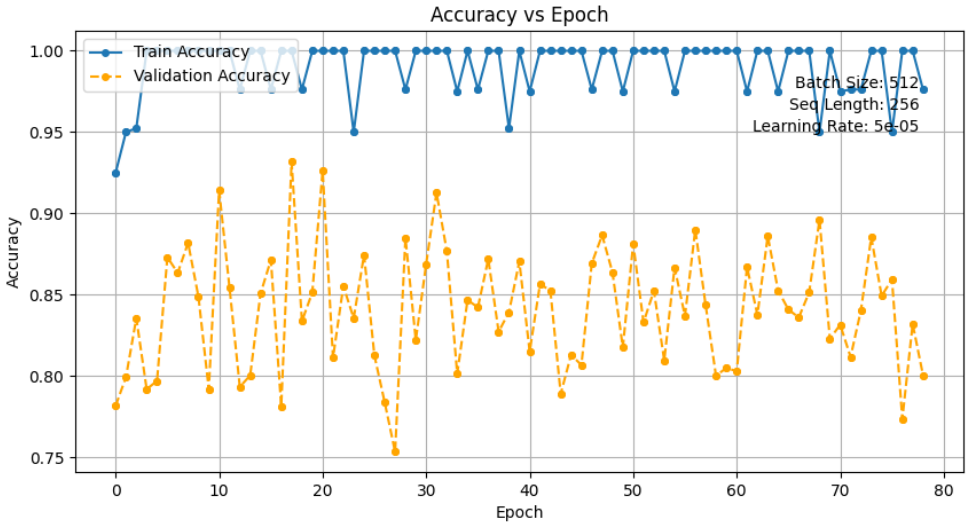
- batch size : 512
  - 에폭 수와 학습량에 따른 조절 수치임
  - 모델과 GPU 환경의 최대 가용 배치 사이즈 수치는 512
- learning rate : 5e-5
- max_seq_length : 256
  - 데이터를 보고 토큰 수 처리 기준은 1024 = 2^10으로 하려함.
  - AI hub에서 가져온 데이터의 글자 수가 대충 700~1000자 사이임
  - 하나의 텐서가 가용할 수 있는 수치 : 300 미만
  - 따라서, 256으로 설정하였음.

- epochs : 79
  - epochs = train data / batch_size = 40000 / 512 ≒ 79

4. 학습 결과

loss_acc_info_batch512_seq256_epoch79_lr5e-05_test07.csv





	Batch Size	Max Seq Length	Epochs	Learning Rate
1	512	256	79	5e-05
2	Epoch	Train Loss	Train Accuracy	Validation Loss,Validat...
3	0	0.15314006805419922	0.925000011920929	0.1659211218357086...
4	0	0.15314006805419922	0.925000011920929	0.1659211218357086...
5	1	0.25144848227500916	0.949999988079071	0.1409371942281723,...
6	1	0.25144848227500916	0.949999988079071	0.1409371942281723,...
7	2	0.1332518756389618	0.9523809552192688	0.1654251217842102,...
8	2	0.1332518756389618	0.9523809552192688	0.1654251217842102,...
9	3	0.023548580706119537	1.0	0.1814982891082763...
10	3	0.023548580706119537	1.0	0.1814982891082763...
156	76	0.0029196147806942...	1.0	0.3445680737495422...
157	77	7.699000707361847e-...	1.0	0.3202364146709442,...
158	77	7.699000707361847e-...	1.0	0.3202364146709442,...
159	78	0.031169572845101357	0.976190447807312	0.456879585981369,0....
160	78	0.031169572845101357	0.976190447807312	0.456879585981369,0....

▼ 실험한 문서들

1. briefing(보도자료)
   {'prediction': '연설문', 'probs': {'뉴스': 0.0205, '보도자료': 0.3573, '역사기록물': 0.0079, '보고서': 0.0063, '회의록': 0.0016, '사설': 0.0019, '간행물': 0.0048, '연설문': 0.5981, '문학': 0.001, '나래이션': 0.0006}}
2. edit(사설)
   {'prediction': '뉴스', 'probs': {'뉴스': 0.959, '보도자료': 0.001, '역사기록물': 0.0004, '보고서': 0.0135, '회의록': 0.0015, '사설': 0.0071, '간행물': 0.0162, '연설문': 0.0006, '문학': 0.0005, '나래이션': 0.0003}}
3. history(역사 기록물)
   {'prediction': '역사기록물', 'probs': {'뉴스': 0.0174, '보도자료': 0.0566, '역사기록물': 0.6604, '보고서': 0.1267, '회의록': 0.0029, '사설': 0.0062, '간행물': 0.1183, '연설문': 0.0028, '문학': 0.0023, '나래이션': 0.0064}}
4. koreabank(연설문)
   {'prediction': '연설문', 'probs': {'뉴스': 0.0003, '보도자료': 0.0115, '역사기록물': 0.0002, '보고서': 0.0007, '회의록': 0.0001, '사설': 0.0002, '간행물': 0.0013, '연설문': 0.985, '문학': 0.0004, '나래이션': 0.0002}}
5. meeting(회의록)
   {'prediction': '보도자료', 'probs': {'뉴스': 0.0014, '보도자료': 0.9946, '역사기록물': 0.001, '보고서': 0.0004, '회의록': 0.0003, '사설': 0.0001, '간행물': 0.0012, '연설문': 0.0005, '문학': 0.0003, '나래이션': 0.0002}}

6. meeting2(회의록)
   {'prediction': '연설문', 'probs': {'뉴스': 0.0099, '보도자료': 0.0567, '역사기록물': 0.0011, '보고서': 0.0691, '회의록': 0.0031, '사설': 0.0182, '간행물': 0.214, '연설문': 0.6171, '문학': 0.0076, '나래이션': 0.0031}}
7. meeting3(보고서) - 있던 데이터 원본 txt
   {'prediction': '보고서', 'probs': {'뉴스': 0.0013, '보도자료': 0.0024, '역사기록물': 0.0261, '보고서': 0.6062, '회의록': 0.0074, '사설': 0.0183, '간행물': 0.0215, '연설문': 0.0495, '문학': 0.0049, '나래이션': 0.2624}}
8. minute1(회의록) - 있던 데이터 원본 txt
   {'prediction': '회의록', 'probs': {'뉴스': 0.0001, '보도자료': 0.0001, '역사기록물': 0.0002, '보고서': 0.0002, '회의록': 0.9988, '사설': 0.0001, '간행물': 0.0001, '연설문': 0.0001, '문학': 0.0002, '나래이션': 0.0001}}
9. news(뉴스)
   {'prediction': '뉴스', 'probs': {'뉴스': 0.9841, '보도자료': 0.0004, '역사기록물': 0.0002, '보고서': 0.0104, '회의록': 0.0003, '사설': 0.0031, '간행물': 0.0005, '연설문': 0.0002, '문학': 0.0002, '나래이션': 0.0004}}
10. news2(사설) - 있던 데이터 원본 txt
    {'prediction': '간행물', 'probs': {'뉴스': 0.0036, '보도자료': 0.0005, '역사기록물': 0.0102, '보고서': 0.1301, '회의록': 0.0019, '사설': 0.2032, '간행물': 0.6316, '연설문': 0.0004, '문학': 0.0134, '나래이션': 0.0052}}
11. test05(사설) - 있던 데이터 원본 txt
    {'prediction': '보고서', 'probs': {'뉴스': 0.0136, '보도자료': 0.0008, '역사기록물': 0.0023, '보고서': 0.8138, '회의록': 0.0009, '사설': 0.0934, '간행물': 0.0718, '연설문': 0.0011, '문학': 0.0005, '나래이션': 0.0018}}
12. ti(문학작품)
    {'prediction': '보고서', 'probs': {'뉴스': 0.0061, '보도자료': 0.001, '역사기록물': 0.0016, '보고서': 0.5984, '회의록': 0.001, '사설': 0.0875, '간행물': 0.0464, '연설문': 0.0013, '문학': 0.2553, '나래이션': 0.0014}}
13. untitled(뉴스)
    {'prediction': '뉴스', 'probs': {'뉴스': 0.9105, '보도자료': 0.0033, '역사기록물': 0.0007, '보고서': 0.0578, '회의록': 0.0006, '사설': 0.0125, '간행물': 0.0104, '연설문': 0.0036, '문학': 0.0003, '나래이션': 0.0003}}