증권 리포트 요약

디지털애널리틱스융합학과 김성주, 이남선

목 차

연구 목적

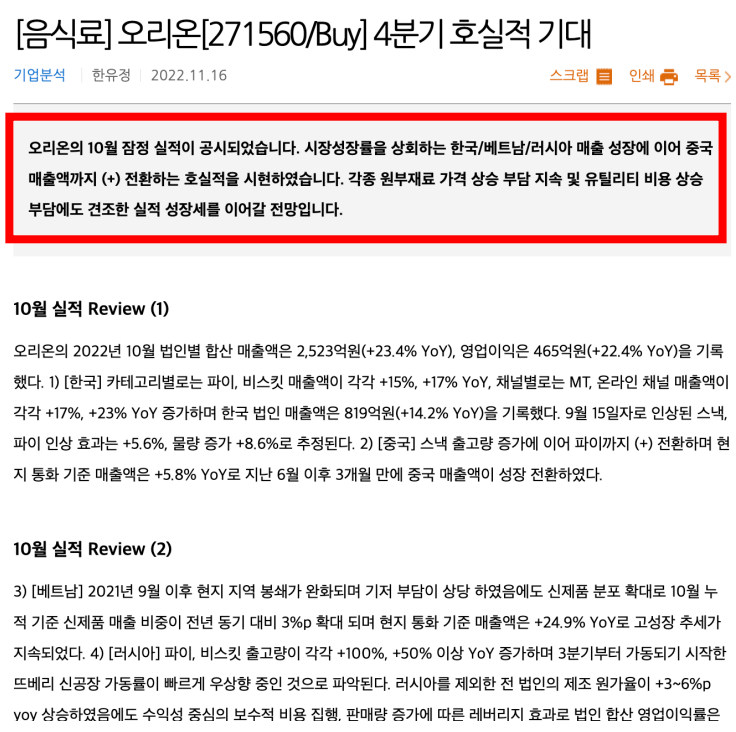
MZ세대를 위한 종목분석 리포트 요약문 제공

증권 리포트 요약문 제공을 위해
금융 도메인에 특화된 KB-ALBERT를 활용한 모델링

목 차

연구 배경

- 증권사 리포트의 경우 전문 용어, 산업 시황, 숫자 등 전문 지식이 없는 일반인이 쉽게 이해하기 어려운 내용으로 구성됨
- 요약문이 제공되는 리포트의 경우에는 이해가 용이하지만, 그렇지 않은 리포트의 경우 핵심을 이해하기에 다소 어려움이 있음



[음식료] 오리온[271560/Buy] 4분기 호실적 기대

기업분석 | 한유정 | 2022.11.16 스크랩 인쇄 목록

오리온의 10월 잠정 실적이 공시되었습니다. 시장성장률을 상회하는 한국/베트남/러시아 매출 성장에 이어 중국 매출액까지 (+) 전환하는 호실적을 시현하였습니다. 각종 원부재료 가격 상승 부담 지속 및 유틸리티 비용 상승 부담에도 견조한 실적 성장세를 이어갈 전망입니다.

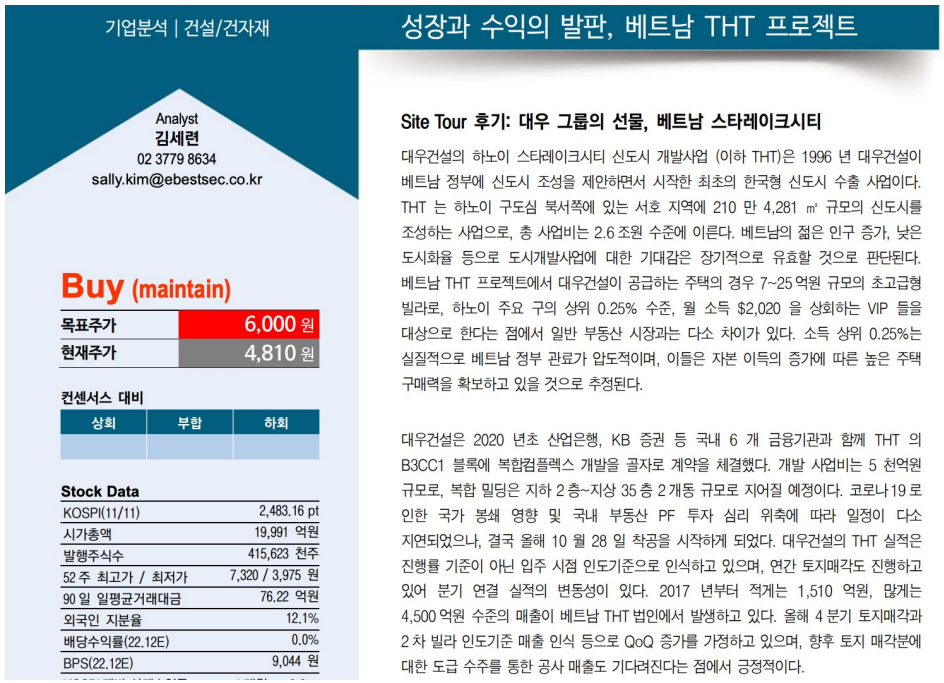
10월 실적 Review (1)

오리온의 2022년 10월 법인별 합산 매출액은 2,523억원(+23.4% YoY), 영업이익은 465억원(+22.4% YoY)을 기록했다. 1) [한국] 카테고리별로는 파이, 비스킷 매출액이 각각 +15%, +17% YoY, 채널별로는 MT, 온라인 채널 매출액이 각각 +17%, +23% YoY 증가하며 한국 법인 매출액은 819억원(+14.2% YoY)을 기록했다. 9월 15일자로 인상된 스낵, 파이 인상 효과는 +5.6%, 물량 증가 +8.6%로 추정된다. 2) [중국] 스낵 출고량 증가에 이어 파이까지 (+) 전환하며 현지 통화 기준 매출액은 +5.8% YoY로 지난 6월 이후 3개월 만에 중국 매출액이 성장 전환하였다.

10월 실적 Review (2)

3) [베트남] 2021년 9월 이후 현지 지역 봉쇄가 완화되며 기저 부담이 상당 하였음에도 신제품 분포 확대로 10월 누적 기준 신제품 매출 비중이 전년 동기 대비 3%p 확대 되며 현지 통화 기준 매출액은 +24.9% YoY로 고성장 추세가 지속되었다. 4) [러시아] 파이, 비스킷 출고량이 각각 +100%, +50% 이상 YoY 증가하며 3분기부터 가동되기 시작한 뜨베리 신공장 가동률이 빠르게 우상향 중인 것으로 파악된다. 러시아를 제외한 전 법인의 제조 원가율이 +3~6%p yoy 상승하였음에도 수익성 중심의 보수적 비용 집행, 판매량 증가에 따른 레버리지 효과로 법인 합산 영업이익률은

〈 요약문이 있는 리포트 〉



기업분석 | 건설/건자재

성장과 수익의 발판, 베트남 THT 프로젝트

Analyst
김세련
02 3779 8634
sally.kim@ebestsec.co.kr

Buy (maintain)

목표주가	6,000 원
현재주가	4,810 원

컨센서스 대비

상회	부합	하회

Stock Data

KOSPI(11/11)	2,483.16 pt
시가총액	19,991 억원
발행주식수	415,623 천주
52 주 최고가 / 최저가	7,320 / 3,975 원
90 일 일평균거래대금	76.22 억원
외국인 지분율	12.1%
배당수익률(22.12E)	0.0%
BPS(22.12E)	9,044 원

Site Tour 후기: 대우 그룹의 선물, 베트남 스타레이크시티

대우건설의 하노이 스타레이크시티 신도시 개발사업 (이하 THT)은 1996 년 대우건설이 베트남 정부에 신도시 조성을 제안하면서 시작한 최초의 한국형 신도시 수출 사업이다. THT 는 하노이 구도심 북서쪽에 있는 서호 지역에 210 만 4,281 ㎡ 규모의 신도시를 조성하는 사업으로, 총 사업비는 2.6 조원 수준에 이른다. 베트남의 젊은 인구 증가, 낮은 도시화율 등으로 도시개발사업에 대한 기대감은 장기적으로 유효할 것으로 판단된다. 베트남 THT 프로젝트에서 대우건설이 공급하는 주택의 경우 7~25 억원 규모의 초고급형 빌라로, 하노이 주요 구의 상위 0.25% 수준, 월 소득 $2,020 을 상회하는 VIP 들을 대상으로 한다는 점에서 일반 부동산 시장과는 다소 차이가 있다. 소득 상위 0.25%는 실질적으로 베트남 정부 관료가 압도적이며, 이들은 자본 이득의 증가에 따른 높은 주택 구매력을 확보하고 있을 것으로 추정된다.

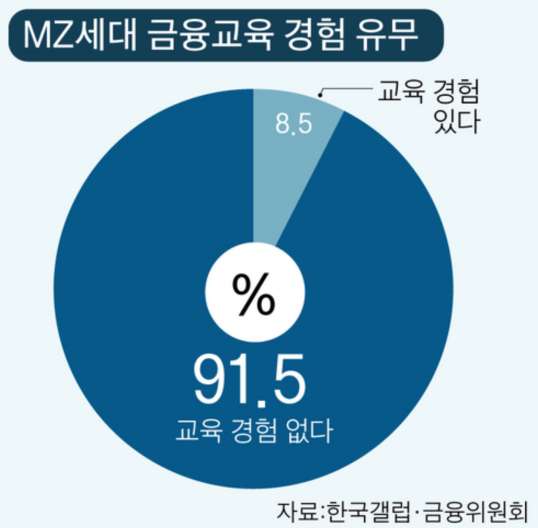
대우건설은 2020 년초 산업은행, KB 증권 등 국내 6 개 금융기관과 함께 THT 의 B3CC1 블록에 복합컴플렉스 개발을 골자로 계약을 체결했다. 개발 사업비는 5 천억원 규모로, 복합 밀딩은 지하 2 층~지상 35 층 2 개동 규모로 지어질 예정이다. 코로나19 로 인한 국가 봉쇄 영향 및 국내 부동산 PF 투자 심리 위축에 따라 일정이 다소 지연되었으나, 결국 올해 10 월 28 일 착공을 시작하게 되었다. 대우건설의 THT 실적은 진행률 기준이 아닌 입주 시점 인도기준으로 인식하고 있으며, 연간 토지매각도 진행하고 있어 분기 연결 실적의 변동성이 있다. 2017 년부터 적게는 1,510 억원, 많게는 4,500 억원 수준의 매출이 베트남 THT 법인에서 발생하고 있다. 올해 4 분기 토지매각과 2 차 빌라 인도기준 매출 인식 등으로 QoQ 증가를 가정하고 있으며, 향후 토지 매각분에 대한 도급 수주를 통한 공사 매출도 기다려진다는 점에서 긍정적이다.

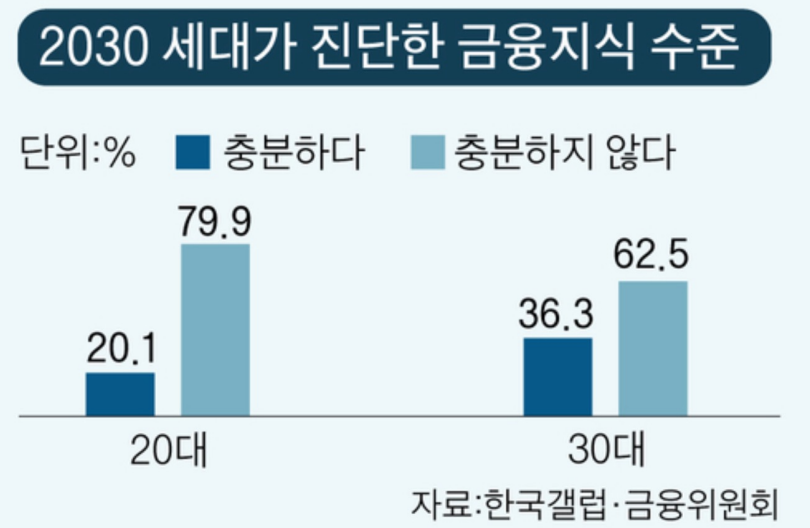
〈 요약문이 없는 리포트 〉

목 차

연구 배경

- MZ세대 금융교육 경험 및 금융지식 수준 낮음
- 하락장일수록 투자공부에 근거한 신중한 투자전략이 필요





무지성 투자가 아니라 정확한 시황분석에 근거한 투자가 필요

[빚의 덫에 걸린 MZ세대] MZ세대, 금융 지식없이 묻지마 단타 투자…정부는 시장 과열 방치해 '빚 폭탄' 키웠다, 중앙선데이, 2022.07.30 https://www.joongang.co.kr/article/25090856
[재테크칼럼] 거대한 변곡점이 온다 – 투자 공부를 해야 하는 이유, 치과신문, 2021.03.25 http://www.dentalnews.or.kr/news/article.html?no=29974

연구 배경

- MZ세대 문해력 낮고 세 줄 이상 읽지 않음
  - '심심한' 사과, '사흘'
  - 숏폼 영상 소비가 많아지면서 긴 글 잘 읽지 않음

효과적인 요약문 필요

"세 줄 이상 안 읽음"...읽는 법을 배우지 않겠다는 뻔뻔한 선언, 톱클래스, 2021.05.25, http://topclass.chosun.com/news/articleView.html?idxno=10789
[뉴스피처] 2틀·4흘로 알다니…몰라도 당당한 디지털 세대, 연합뉴스, 2020.07.31, https://www.yna.co.kr/view/AKR20200730147000797

목 차

데이터 소개

1. 증권사 리포트

- 학습 데이터 : 요약문이 있는 증권사 리포트 (약 8천 건)
  - 키움증권 - 2012년 11월 이후 리포트 약 4504건
    - PDF 다운로드 후 pdf2text 작업 진행
    - 엑셀 파일로 요약문과 본문 분리하는 라벨링 작업 진행
  - 한화투자증권 - 2016년 1월 이후 리포트 약 3340건
    - 크롤링으로 요약문과 본문 따로 수집 후 엑셀 파일 저장

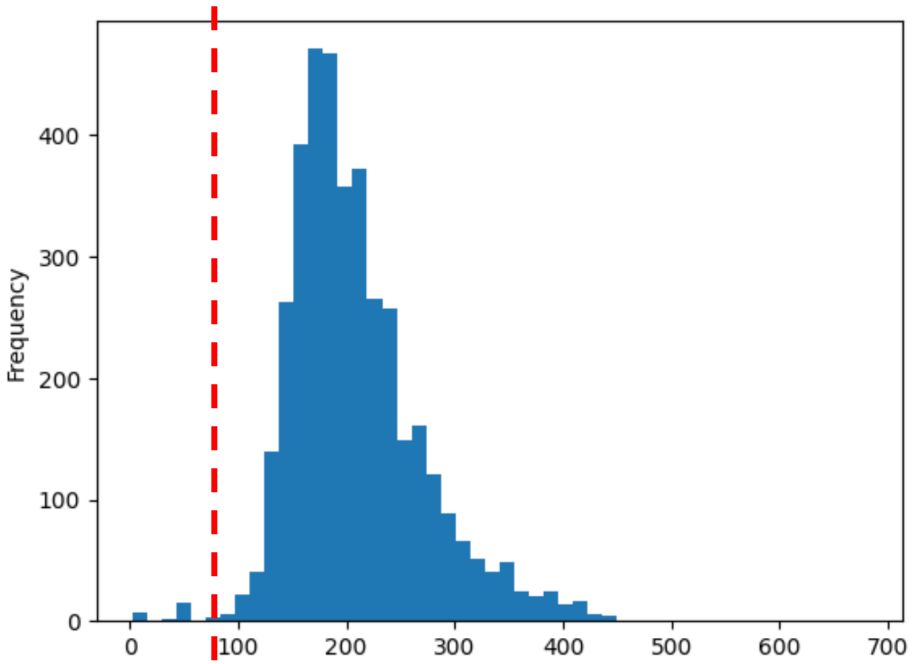
2. AI-HUB

- 요약문 및 레포트 생성 데이터 (동일 데이터 유형) - '데이터 종류 - 보고서'만 사용 (9000건)
- 문서 요약 텍스트 데이터 (동일 도메인) - 신문기사 30만 건 중 '경제지(매체유형)'만 사용(2만2342건)

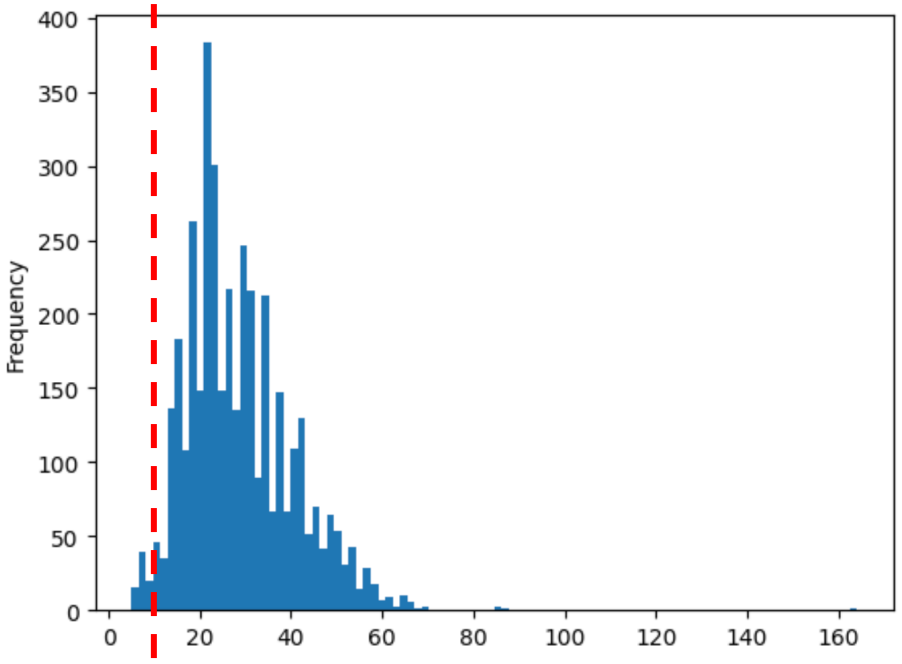
목 차

데이터 소개

- 데이터 길이 처리
  - 분포 길이 확인 후 본문 길이 80 미만, 요약문 길이 10 미만 데이터 삭제



본문 길이 분포



요약문 길이 분포

- 최종 38,272 데이터를 Train : Valid : Test = 8 : 1 : 1로 split

KB-ALBERT

- KB국민은행에서 제공하는 경제/금융 도메인에 특화된 한국어 ALBERT 언어모델
- 모델 파일은 비영리목적으로 요청 시에만 개별 제공

도메인	Task	BERT Base Multi-Lingual	KB-ALBERT
일반	감성분류 (Naver)	0.888	0.91
일반	MRC(KorQuAD 1.0)	0.87	0.90
금융	MRC(자체)	0.77	0.89

*ALBERT는 BERT 계열의 사전학습 모형으로 BERT 모델을 경량화 하고
Next Sentence Prediction 대신 Sentence Order Prediction 학습 목표를 가짐

KB-ALBERT

- 학습 데이터셋 :
  KB국민은행이 보유하고 있는 약 1억건 이상의 텍스트 데이터,
  한 영역에 치우치지 않도록 일반 도메인의 학습 데이터도 많은 양 사용
  - 일반 도메인 텍스트(위키 + 뉴스 등) : 약 25GB
  - 금융 도메인 텍스트(경제/금융 특화 뉴스 + 리포트 등) : 약 15GB
  - ➢ version 2에서는 100GB로 늘려서 학습

- Tokenizer :
  - 음절단위 한글 토크나이저 사용 (BertWordPieceTokenizer에서 음절만 있는 형태와 비슷)

- 모델 Architecture :

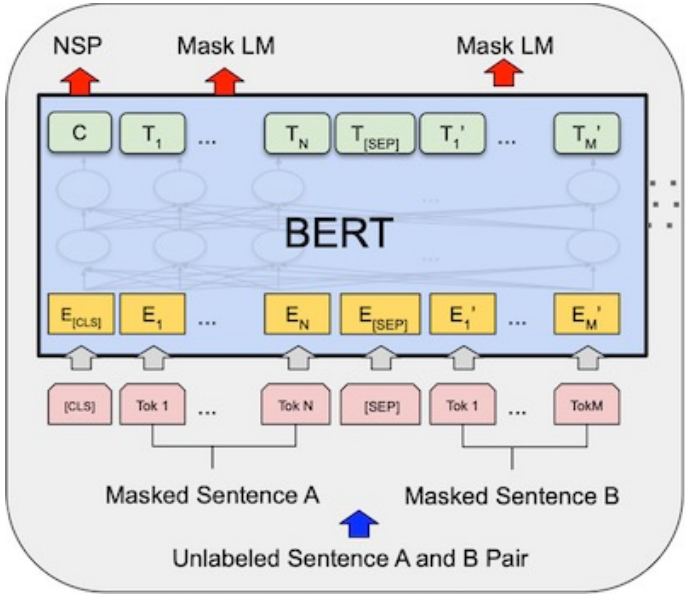
max_seq_length	embedding_size	hidden_size	num_hidden_layers	vocab_size
512	128	768	12	9,607

KoBART 모델

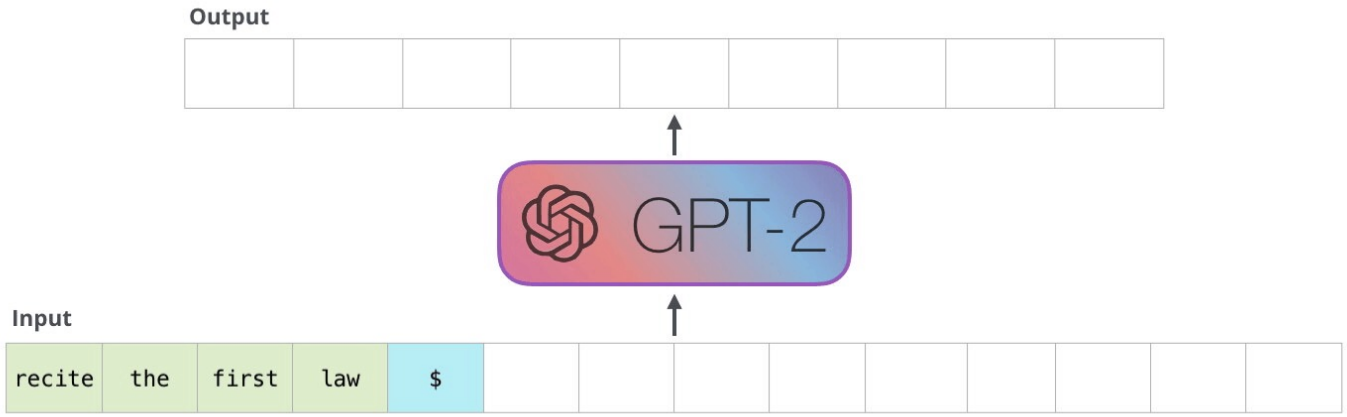
- KoBART : BART 모델 구조와 task 그대로 한국어에 특화 (SKT)

<BART>

- Bidirectional과 Auto-Regressive Transformer를 합친 denoising autoencoder
  - 다양한 downstream 태스크에서도 잘 작동



BERT : Bidirectional

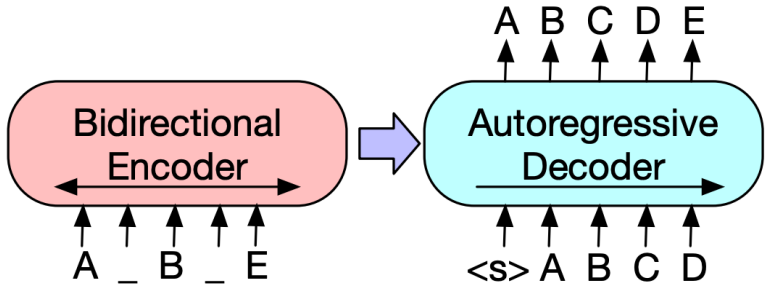


GPT : Auto-Regressive

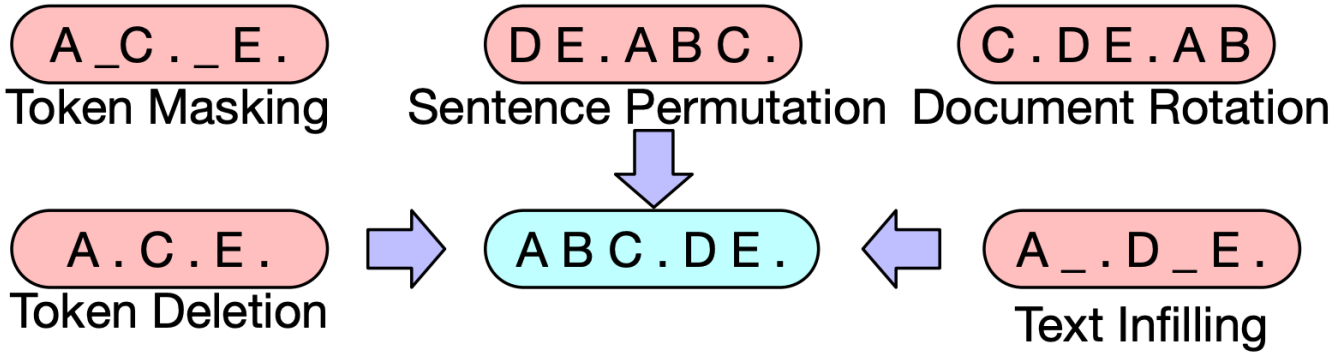
KoBART 모델

<BART>

- denoising autoencoder



- 새로운 Pre-training objective로 학습 : 오염된 텍스트를 복원



- encoder, decoder 모두 6개의 층으로 구성 + ReLU 대신 GeLU 활성화 함수 사용
- encoder의 최종 은닉층에 대해 cross-attention
  additional feed-forward network는 사용x

KoBART 모델

- 학습 데이터셋 :
  한국어 위키 백과 5백만 개 문장 이외에, 뉴스, 책, 모두의 말뭉치 v1.0,
  청와대 국민청원 등의 다양한 데이터가 모델 학습에 사용됨

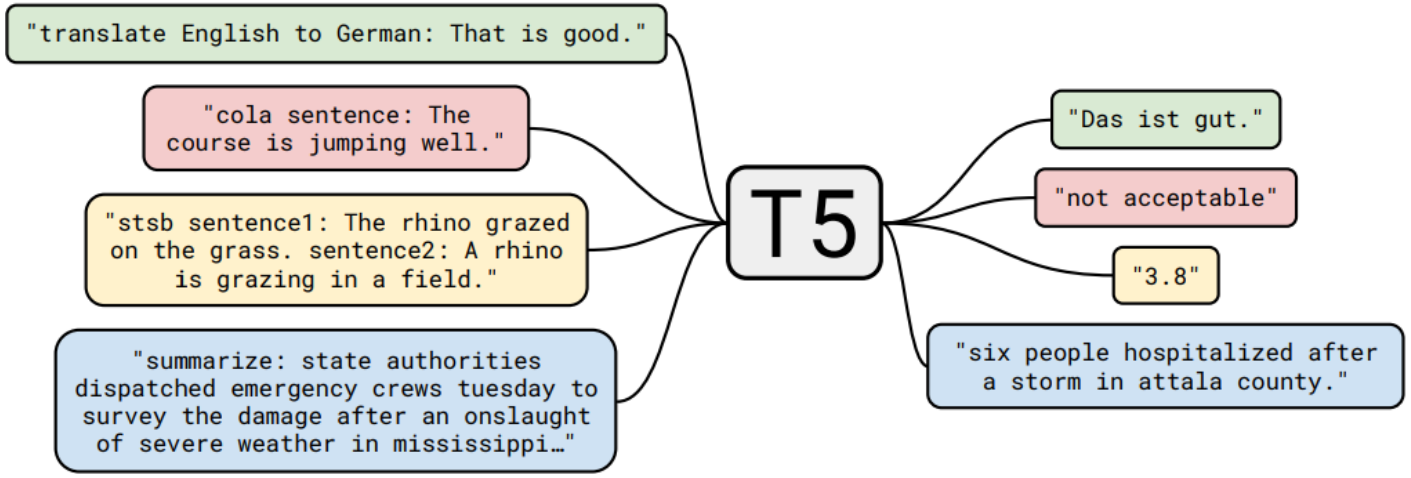
- 모델 Architecture :

# of params	embedding_size	hidden_size	num_layers	vocab_size
124M	Encoder: 768 Decoder: 768	3,072	Encoder: 6 Decoder: 6	30,000

목 차

KoT5

- T5(text-to-text-transfer-transformer)는 입력과 출력이 항상 텍스트 문자열인 Text to Text 프레임워크를 사용하여 모든 NLP Task들을 일반화
- 학습 데이터셋 : 42GB의 한국어 말뭉치 사용 (위키 백과 및 신문기사 등 23개 종류)



- 모델 Architecture :

# of params	embedding_size	hidden_size	num_layers	vocab_size
250M	512	2,048	Encoder: 6 Decoder: 6	32,128

실험 내용

- KB-ALBERT의 tokenizer 단어사전 비정상 ⟶ 모델 사용불가 판단

5250	툰	3120	配	6800	##꼽	8116	##K
5251	툴	3121	酒	6801	##●	8117	##И
5252	툼	3122	酬	6802	##갑	8118	##E
5253	툽	3123	酷	6803	##뒷	8119	##R
5254	퉁	3124	酸	6804	##쪽	8120	##₩
5255	퉈	3125	醉	6805	##넬	8121	##뒀
5256	퉜	3126	醒	6806	##뷔	8122	##칰
5257	퉤	3127	醜	6807	##🍺	8123	##몹
5258	튀	3128	醫	6808	##극	8124	##👅
5259	튄	3129	醬	6809	##볼	8125	##맸
5260	튈	3130	采	6810	##²	8126	##쌉
5261	튕	3131	釋	6811	##○	8127	##넋
5262	튜	3132	里	6812	##ㅉ	8128	##ㅀ
5263	튠	3133	重	6813	##ㄴ	8129	##↑
5264	튤	3134	野	6814	##ㅎ	8130	##갯
5265	튬	3135	量	6815	##먼	8131	##👴
5266	튭	3136	金	6816	##y	8132	##😔

KB-Albert Tokenizer 단어사전

```
[('_총괄', 20090),
 ('_면접', 20091),
 ('_지연', 20092),
 ('_분쟁', 20093),
 ('_벤처', 20094),
 ('_남부', 20095),
 ('_소외', 20096),
 ('_반도체', 20097),
 ('ind', 20098),
 ('_요구했다.\n', 20099),
 ('_지방선거', 20100),
 ('_힘입', 20101),
 ('_이적', 20102),
 ('숙한', 20103),
 ('_무엇이', 20104),
 ('_벌이고', 20105),
 ('_대만', 20106),
 ('_검찰이', 20107),
 ('십시오', 20108),
 ('_내다봤다.\n', 20109),
 ('있게', 20110),
 ('_국가의', 20111),
```

KoBART Tokenizer 단어사전

실험 내용

- KoBART와 KoT5 두 가지 모델만 가지고 요약 성능 비교
- KoBART 모델은 사전학습 모델의 파라미터를 그대로 사용한 버전과 fine-tuning한 버전 두 가지로 학습
- KoT5 모델은 fine-tuning한 버전으로 epoch 2로 학습

Hyperparameter	KoBART (feature-based)	KoBART (fine-tuning)	KoT5 (fine-tuning)
epoch	50	50	2
batch size	4	4	4
max length	512	512	512
learning rate	3e-5	3e-5	1e-4

목 차

실험 내용

Golden Summary

메르스 효과로 인한 **중국 노선 부진**이 성수기 내내 이어졌지만 **3분기 대한항공은 2,503억원의 양호한 영업이익**을 기록했을 전망입니다. 다만 **원화 가치 하락에 따른 외화환산 손실로 순적자**는 불가피해 보입니다. 최근 LCC와의 가격 경쟁이 심화되고 있는 것은 사실이지만 **LCC와는 차별적인 최신 대형기 도입**으로 **중국 노선 회복**과 **일본 노선의 계속되는 강세** 효과, 그리고 **환승 고객 유치**가 기대됩니다.

1. KoBART (Fine-tuned)

대한항공은 3분기 영업이익 2,503억원(+4.0%)를 기록했을 것으로 전망합니다. 내국인 출국자 회복세가 예상보다 빠르게 나타나고 있고, 국제 유가도 정상화되고 있기 때문입니다. 또한 **신형 대형기 투입**을 통해 장거리 **환승 수요**와 서비스에 민감한 **상용 수요를 확보**하고 있어 정부의 환승가능시간 확대 정책의 수혜도 기대할 수 있을 것으로 판단합니다.

2. KoBART (Feature-based)

지난 3분기 대한항공은 IFRS 연결 기준 매출액 2조9,686억원(YoY 기준 -6.2%), **영업이익 2,503억원**(+4.0%)을 기록했을 것으로 전망하며 **메르스 영향**은 내국인 출국자의 경우 상대적으로 빠른 회복세가 나타났지만 **중국인입국자**의 경우는 **감편된 항공편이 정상화된 이후에 점차 회복**되고 있는 상황이어서 3분기 실적에는 크게 기여하지 못했을 것으로 추정된다.

3. KoT5 (Fine-tuned)

대한항공의 3분기 영업이익은 2,503억원으로 시장 기대치에 부합했을 전망이다. **메르스 효과로 인한 중국 노선 부진**이 성수기 내내 이어졌지만 **원화 가치 하락에 따른 외화환산 손실로 순적자**를 기록할 것으로 보인다.

ROUGE score

- Recall-Oriented Understudy for Gisting Evaluation
- 텍스트 요약, 기계 번역 등 자연어 생성 모델의 성능을 평가하기 위한 지표
- 모델이 생성한 요약과 사람이 만들어 놓은 참조본을 대조해 성능 점수 계산

모델이 생성한 요약본	참조 요약본(정답 데이터)
The cat was found under the bed	The cat was under the bed

겹치는 단어 수 = 6

- Recall과 Precision 계산
  - Recall: **참조 요약본**을 구성하는 단어 중 몇 개의 단어가 **모델 요약본**의 단어들과 겹치는지
  - Precision: 모델이 생성한 **요약본** 중 **참조 요약본**과 겹치는 단어들이 얼마나 많이 존재하는지
- 정확한 성능 평가를 위해 Recall과 Precision 계산 후, F1 score 사용
- ROUGE-N, ROUGE-S, ROUGE-L 다양한 지표 존재

ROUGE score

- ROUGE-N : 문장 간 중복되는 n-gram 수 비교하는 지표
  - ROUGE-1 : 모델 요약본과 참조 요약본 간 겹치는 unigram의 수
  - ROUGE-2 : 모델 요약본과 참조 요약본 간 겹치는 bigram의 수

- ROUGE-L : LCS(Longest Common Sequence)기법을 이용해 최장 길이로 매칭되는 문자열 측정
  - N-gram과 달리 순서나 위치 관계를 고려한 알고리즘
  - ROUGE-2와 같이 연속적인 매칭을 요구하지 않고 문자열 내에서 발생하는 매칭 측정으로 보다 유연한 성능 비교 가능
  - Recall : LCS 길이 / 참조 요약본 N-gram의 수
  - Precision : LCS 길이 / 모델 요약본 N-gram의 수

- ROUGE-S : 특정 Window size가 주어졌을 때, Window size 내에 위치하는 단어쌍들을 묶어 해당 단어쌍들이 얼마나 중복되게 나타나는 지 측정

모델 성능 비교

- 평가 지표 확인 결과, KoBART fine-tuning 모델이 가장 성능이 높은 것으로 나타남
- KoT5 모델은 epoch 2로 작게 fine-tuning 했음에도 epoch 50의 feature-based KoBART와 유사

	ROUGE 1			ROUGE - L		
	Recall	Precision	F1	Recall	Precision	F1
KoBART(finetuing, epochs 50)	0.2961	0.2839	0.2802	0.2741	0.2633	0.2597
KoBART (feature-based, epochs 50)	0.2570	0.2861	0.2573	0.2398	0.2675	0.2403
KoT5(epochs 2)	0.2873	0.2873	0.2769	0.2629	0.2785	0.2683

결론

- 금융 도메인에 특화된 KB-ALBERT 모델을 사용하여 증권사 리포트를 요약하려는 것이 최초 아이디어였으나 토크나이저 문제로 사용하지 못하게 된 것에 대한 아쉬움

- text summarization task에 높은 성능을 보이는 KoBART와 KoT5 사전학습 모형을 활용해 금융 도메인 특화 미세조정 모델을 시도했다는 의의
  - 단, KoT5 모델은 epoch 증가시켜 실험할 필요

- 영어와 다른 한국어 특징을 반영할 수 있는 평가 지표인 RDASS로 성능을 비교하면 좀 더 합리적인 성능 비교가 가능했을 것으로 생각됨

Reference

- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140), 1-67.
- Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74 - 81, Barcelona, Spain. Association for Computational Linguistics.
- Lee, D., Shin, M., Whang, T., Cho, S., Ko, B., Lee, D., ... & Jo, J. (2020). Reference and document aware semantic evaluation methods for Korean language summarization. arXiv preprint arXiv:2005.03510.
- KoBART-Summarization github code https://github.com/seujung/KoBART-summarization
- KB-ALBERT github code https://github.com/KB-AI-Research/KB-ALBERT
- T5 github code https://github.com/wisenut-research/KoT5
- BART paper review https://chloelab.tistory.com/34
- BART paper review https://kubig-2021-2.tistory.com/m/50
- ALBERT paper review https://jeonsworld.github.io/NLP/albert/
- T5 paper review https://velog.io/@mooncy0421/Paper-Review-T5-Exploring-the-Limits-of-Transfer-Learning-with-a-Unified-Text-to-Text-Transformer

Github

- 본 프로젝트 관련 코드 및 증권 리포트 csv파일 github 업로드

  https://github.com/sunnie720/stockReport_summarization

감사합니다