한국어 뉴스 헤드라인의 토픽 분류에 대한 실증적 연구

박제윤[1○*], 김민규[2*], 오예림[3*], 이상원[4*], 민지웅[5*], 오영대[1†]
[1]엘솔루, [2]동국대학교, [3]한국외국어대학교, [4]인천대학교, [5]애자일소다
{jeiyoon.park,youngdae.oh}@llsollu.com, kmg2933@dgu.ac.kr,
bora7474@hufs.ac.kr, leo503801@inu.ac.kr, jayden@agilesoda.ai

An Empirical Study of Topic Classification for Korean Newspaper Headlines

Jeiyoon Park[1○*], Mingyu Kim[2*], Yerim Oh[3*], Sangwon Lee[4*], Jiung Min[5*], Youngdae Oh[1†]
[1]LLSOLLU, [2]Dongguk University, [3]Hankuk University of Foreign Studies, [4]Incheon University, [5]Agilesoda

요 약

좋은 자연어 이해 시스템은 인간과 같이 텍스트에서 단순히 단어나 문장의 형태를 인식하는 것 뿐만 아니라 실제로 그 글이 의미하는 바를 정확하게 추론할 수 있어야 한다. 이 논문에서 우리는 뉴스 헤드라인으로 뉴스의 토픽을 분류하는 open benchmark인 KLUE(Korean Language Understanding Evaluation)에 대하여 기존에 비교 실험이 진행되지 않은 시중에 공개된 다양한 한국어 라지스케일 모델들의 성능을 비교하고 결과에 대한 원인을 실증적으로 분석하려고 한다. KoBERT, KoBART, KoELECTRA, 그리고 KcELECTRA 총 네가지 베이스라인 모델들을 주어진 뉴스 헤드라인을 일곱가지 클래스로 분류하는 KLUE-TC benchmark에 대해 실험한 결과 KoBERT가 86.7 accuracy로 가장 좋은 성능을 보여주었다.

주제어: Topic classification, Large-scale language model

1. 서론

자연어 이해(Natural Language Understanding, NLU) [1]란 인간의 언어표현을 기계가 의미를 이해할 수 있도록 변환시키는 것을 의미한다. 좋은 NLU 시스템은 인간과 같이 글에서 단순히 단어나 문장의 형태를 인식하는 것 뿐만 아니라 실제로 그 글이 의미하는 바를 추론할 수 있어야 한다. 예를 들어, 그림 1에서 *"인공지능 포커서도 프로 4명에 압승...20억 원 이상 따"*라는 뉴스 헤드라인이 있을때, 좋은 NLU 시스템은 사람처럼 이 뉴스가 *IT/Science*와 관련된 뉴스를 다룰 것이라고 예측할 수 있어야 한다.

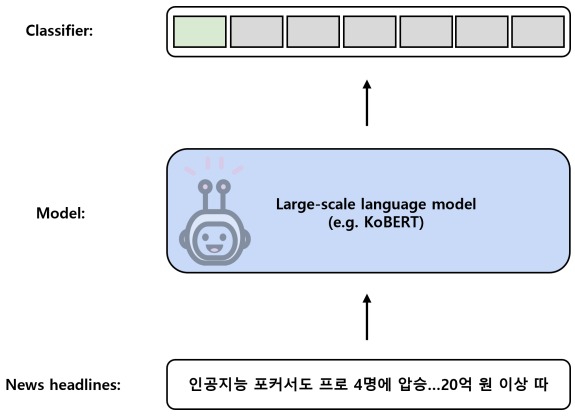

그림 1. 한국어 large-scale language model을 이용한 뉴스 헤드라인의 토픽 분류 예시. 좋은 토픽 분류 시스템 단순히 문장의 구조나 형태만 인식하는게 아니라, 실제 사람처럼 문장이 의미하는 바를 명확하게 추론할 수 있어야 한다.

딥러닝 기반 모델의 발전에 따라 최근 NLU에 관한 연구는 전통적인 규칙 기반의 머신러닝 기법을 벗어나 토픽 분류(Topic Classification) [2], 개체명 인식(Named Entity Recognition) [3], QA(Question Answering) [4], 기계 번역(Machine Translation) [5], 문서 요약(Text Summarization) [6], 기계 독해(Machine Reading Comprehension) [7] 등 다양한 분야에 적용되고 있다.

이 중에서 토픽 분류(Topic Classification, TC)는 text snippet이 주어졌을때, snippet에 내포된 토픽을 예측하는 task이다 [2, 8]. 기존에는 TC task를 위한 한국어 표준 benckmark 가 존재하지 않아서 여러 모델 사이의 성능을 체계적으로 비교하는 것은 어려움이 있었지만, 최근 공개된 KLUE [9]의 KLUE-TC benchmark 덕분에 이 문제가 해소되었다.

한국어 기반의 토픽 분류 모델은 감성 분류(Sentiment Classification), 의도 분류(Intent Classification) 등 다양한 분야에 적용이 가능하고 중요하게 다뤄지는 NLU task인 만큼 다양하고 활발한 연구가 진행되어야 하는 분야이다.

좋은 성능의 분류 모델을 만들기 위해 가장 많이 활용하는 방법은 Transformer [10] 기반의 사전학습 언어 모델이다. Transformer의 encoder를 활용한 모델인 BERT [11] 모델이 등장한 이후 분류 task를 수행하기 위한 방법으로 대용량의 말뭉치로 pretrain된 모델에 fine-tuning을 적용하는 방법이 일반화되었다.

이후 BERT에 더 많은 데이터와 Dynamic Masking을 적용

*These authors contributed equally.
†Corresponding author.

해 성능을 끌어올린 RoBERTa [12], Transformer의 decoder를 활용하여 자연어 생성에서 좋은 성능을 보여주는 GPT-n [13, 14, 15], 다섯가지 noise로 손상시킨 데이터로 pretrain시켜 NLU와 NLG 모델에 함께 활용할 수 있는 BART [16] 등 여러 모델 등이 등장했다.

이들 중 한국어로 사전학습되어 시중에 공개된 large-scale language model은 KorBERT[1], KoBERT[2], KoBART[3], KoELECTRA[4], 그리고 KcELECTRA[5] 등이 있다.

한국어 기반의 large-scale language model은 최근 네이버에서 공개한 HyperClova [17]까지 많은 발전을 거듭하고 있다.

이 논문은 한국어 토픽 분류 표준 벤치마크인 KLUE-TC benchmark에 대하여 시중에 공개되어 있지만 아직 성능 비교 실험이 진행되지 않은 한국어 large-scale language model들의 실증적인 실험과 비교, 그리고 분석을 진행하였다. KLUE의 task-classification 데이터셋인 KLUE-TC를 활용해 KoBERT, KoBART, KoELECTRA, KcELECTRA의 성능을 비교해본 결과 KoBERT가 86.7%로 pretraining 방법과 corpus 구성 방법을 개선시킨 다른 모델들에 비해 가장 좋은 성능을 보였다.

2. Language Model

2.1 KoBERT

BERT [11]는 2018년에 구글이 공개한 pretrained model로 다양한 NLP task에서 우수한 성능을 보이는 모델이다. BERT 모델의 동작 과정은 크게 Data Input, Pretraining, Fine-Tuning 3가지 단계로 이루어진다.

(1) Data Input: Word Piece 토크나이저를 사용해서 단어보다 작은 단위로 쪼개는 Token Embeddings, 두 문장을 구분해서 문장을 학습하는 Segment Embeddings, Transformer[10]에서 단어의 위치를 표현하는 Position Embeddings 세 가지를 더해 모델의 Input으로 사용한다. (2) Pretraining: BERT의 사전 훈련은 Masked Language Model (MLM)과 Next Sentence Prediction (NSP)으로 진행된다. MLM은 입력 데이터의 15%의 단어를 무작위로 마스킹한 후 인공 신경망이 마스킹된 단어를 예측하게 한다. NSP는 두 개의 문장이 연결되는 문장인지를 맞추는 방식으로 훈련을 진행한다. (3) Fine-Tuning: 해결해야 할 Task의 데이터로 모델을 추가로 학습시키는 단계이다.

본 논문에서는 대용량 한국어 wiki corpus로 학습한 SKT의 KoBERT를 사용하였다.

[1] https://aiopen.etri.re.kr/service_dataset.php
[2] https://github.com/SKTBrain/KoBERT
[3] https://github.com/SKT-AI/KoBART
[4] https://github.com/monologg/KoELECTRA
[5] https://github.com/Beomi/KcELECTRA

2.2 KoBART

BART [16]은 2019년 Facebook이 공개한 모델로 Denoising Autoencoder[18] 방식으로 학습된다. Noise가 추가된 데이터에서 Noise를 제거하면서 기존의 데이터를 추출하는 방법이다. BART는 표준 sequence-to-sequence Transformer 구조를 사용하며, 기존 활성화 함수 ReLU를 GeLUs로 변경했다.

Pretraining: BART의 사전 훈련을 위한 입력 데이터는 (1) Token Masking (2) Token Deletion (3) Token Infilling (4) Sentence Permutation (5) Document Rotation 다섯 가지의 Noising 기법을 제시하고 있다. 본 논문에서 실험에 사용한 한국어 BART는 Text Infilling 기법을 사용했다. Text Infilling 기법은 Poisson distribution으로 부터 얻은 span 길이를 이용해 text span을 sampling한다. 각각의 span은 단일 [MASK] Token으로 대체된다. span의 길이가 0인 경우 [MASK] Token이 삽입된다.

Fine-tuning: encoder와 decoder에 동일한 입력을 넣고 마지막 decoder token에 마지막 hidden state를 다중 클래스 분류기의 입력으로 사용하는 Sequence Classification을 이용했다.

2.3 KoELECTRA

기존 BERT [11]를 포함한 많은 모델들은 입력을 mask token으로 치환하고 이를 다시 치환 전의 원본 token으로 복원하는 방식으로 pretraining을 한다. 하지만 이런 방법을 사용했을 때 많은 계산량이 필요하다는 단점이 있다.

하지만 ELECTRA [19]모델은 Replaced Token Detection (RTD)를 사용해 기존 모델들보다 빠르고 효과적으로 학습한다. RTD는 generator를 사용해 실제 입력의 일부를 가짜 토큰으로 바꾸고, 각 토큰이 실제 입력에 있는 진짜 토큰인지 generator가 생성해낸 가짜 토큰인지 discriminator가 맞히는 이진 분류 문제이다.

기존 BERT에서 사용되는 Masked Language Modeling (MLM) 태스크의 경우 입력 시퀀스 토큰 중 약 15%만 사용해 학습하지만 RTD 태스크의 경우 입력의 15%가 아니라 모든 토큰에 대해 학습을 진행하기 때문에 학습 속도가 빠르다. 여러 실험 결과에서도 ELECTRA가 BERT보다 효율적으로 학습한다는 결과를 얻었다.

KoELECTRA는 ELECTRA 모델을 가지고 34GB의 한국어 text로 pretraining한 모델로 KoELECTRA-Base-v3 모델의 경우 NSMC Task에서 정확도가 90.63으로 KoBERT의 89.59를 능가하는 성능이 나왔다.

KoELECTRA는 다른 모델에 비해 학습 속도가 빠르고 한국어 분류 Task에서 BERT를 넘어서는 성능을 보여주고 있다. 우리는 이러한 점에 영감을 받아 한국어 뉴스 헤드라인 토픽 분류 문제에 KoELECTRA를 활용하였다.

표 1. 한국어 뉴스 헤드라인의 토픽 분류 Benchmark (KLUE-TC).

Topic	Label	Headline Examples	#Training	#Test
IT/Science	0	내년 첫 5G폰 평균가 80만원 육박... 2023년 60만원대 ↓	4,824	
Economy	1	코스피 기관·개인 매수에 반등... 코스닥 1%대 상승종합	6,222	
Society	2	뜨거운 감자 포털 규제... 필요성·시행안 등 논쟁 가열종합	7,362	
Culture	3	웹툰 나이트에 참가한 작가들	5,933	
World	4	총기폭력 더 방치 못해... 美 초강력 총기규제안 도입종합	7,629	
Sports	5	아시안피스컵 남북 男배구 열전 펼쳐... 북한팀 32 승종합	6,933	
Politics	6	국회사무처 美매사추세츠大·뉴욕시립대 퀸즈칼리지와 MOU	6,751	
Total			**45,654**	**9,131**

2.4 KcELECTRA

KoELECTRA를 포함해 기존에 공개된 한국어 Transformer 계열 모델들은 대부분 한국어 위키, 뉴스 기사, 책 등 잘 정제된 데이터를 기반으로 학습한 모델이라 구어체 특징에 신조어나 오탈자 등 공식적인 글쓰기에서 나타나지 않는 표현들이 빈번하게 등장하는 Task에서 성능이 떨어진다.

KcELECTRA는 네이버 뉴스의 댓글과 대댓글 약 17.3GB, 1억8천만개 이상의 문장을 수집해, 토크나이저와 ELECTRA 모델을 처음부터 Pretraining한 ELECTRA모델로, 위와 같은 Task에서 성능이 뛰어나다.

NSMC Task에서 KoELECTRA가 90.63인데 반해 KcELECTRA는 91.71로 높은 성능을 보이고 있다. 우리는 이러한 점에 영감을 받아 한국어 뉴스 헤드라인 토픽 분류 문제에서도 높은 성능을 기대해 KcELECTRA를 활용하였다.

3. 실험

3.1 Task

우리는 baseline들의 표준적인 성능을 비교하기 위해 KLUE-TC benchmark를 사용하였다. KLUE-TC는 Yonhap News Agency에서 2016년 1월부터 2020년 12월까지 수집한 뉴스 헤드라인들로 구성되어 있고 각 헤드라인들은 IT/Science, Economy, Society, Culture, World, Sports, Politics 총 일곱개의 클래스로 분류된다. 표 1 은 KLUE-TC task의 각 토픽별 예시와 학습과 평가에 사용된 데이터 수를 나타낸다. 또한, 그림 2 은 KLUE-TC task의 각 topic별 학습 데이터의 비율을 보여준다.

베이스라인은 기존에 표준 벤치마크 성능 비교가 되지 않은 시중에 오픈된 네가지 한국어 large-scale language model인 KoBERT, KoBART, KoELECTRA, KcELECTRA를 사용하였다. Evaluation metric으로는 accuracy를 사용하였다.

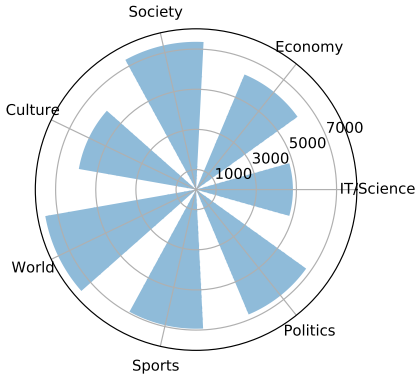

그림 2. KLUE-TC benchmark의 각 토픽별 데이터 수의 시각화.

Method	Accuracy
KoBART	85.62
KoELECTRA	84
KcELECTRA	84.57
KoBERT	**86.7**

표 2. KLUE-TC benchmark에 대한 KoBERT, KoBART, KoELECTRA, KcELECTRA의 실험 결과.

3.2 결과

표 2는 KLUE-TC benchmark에 대한 네가지 베이스라인 모델들의 accuracy 성능을 나타낸다. 이 실험에서 KoBERT의 성능이 86.7로 가장 높았고 KoBART는 85.62, KcELECTRA는 84.57를 각각 기록하였다. KoELECTRA는 가장 낮은 84를

Method	Accuracy
KoBART	90.24
KoELECTRA	90.63
KcELECTRA	**91.54**
KoBERT	89.59

표 3. NSMC benchmark에 대한 KoBERT, KoBART, KoELECTRA, KcELECTRA의 실험 결과. 각 모델의 성능은 기존에 공개된 NSMC benchmark 실험들의 protocol을 그대로 따라서 진행하였다.

Hyperparameter	Value
Max length	64
Batch size	32
Warmup ratio	0.1
Epochs	4
Max grad norm	1
Log interval	200
Learning rate	4e-5

표 4. 실험에 사용한 KoBERT 모델의 hyperparameter.

기록하였다.

표 3에서 NSMC task 성능을 보면 KoBERT가 89.59로 가장 낮은 점수를 기록하였고, KoBART가 90.24, KoELECTRA가 90.63을 각각 얻었고, KcELECTRA가 91.54로 가장 높은 성능을 보인다. NSMC benchmark의 결과를 KLUE-TC benchmark 실험과 비교했을때 정반대의 결과가 나오게 되는데 이를 통해 KLUE-TC benchmark에 대해서는 모델 사이즈는 비슷하지만 한국어 wiki를 사용하여 pretraining한 KoBERT 모델이 한국어 위키, 뉴스, 나무위키, 그리고 모두의 말뭉치 제공 데이터를 사용하여 학습시킨 KoELECTRA , 그리고 네이버 뉴스 댓글과 대댓글 데이터를 사용하여 학습한 KcELECTRA 보다 성능이 좋게 나오는 것을 알 수 있었다.

또한 기존의 다섯가지 noise를 추가하여 학습한 BART와 다르게 Text infilling 한가지 노이즈만 추가하여 pretraining 한 KoBART 역시 KoBERT 보다 성능이 낮게 나왔다. 이는 KLUE-TC benchmark task에서 Text infilling noise만 추가하여 pretraining 하는게 성능에 크게 영향을 주지 않았음을 의미한다.

실험에 사용된 KoBERT 모델의 hyperparameter는 표 4와 같다.

4. 결론

이 논문에서 우리는 인간과 같이 글이 의미하는 바를 정확하게 추론할 수 있는 좋은 자연어 이해 모델은 어떤 학습 조건을 필요로 하는가에 대한 의문에서 시작하여 기존에 표준 벤치마크 실험이 진행되지 않은 한국어 large-scale language model들에 대해 뉴스 헤드라인을 분류하는 KLUE-TC benchmark 실험을 진행하였다.

실험 결과 네가지 베이스라인 중 KoBERT의 accuracy 성능이 86.7으로 가장 높게 나왔다. 이를 통해 우리는 기존에 오픈된 한국어 모델들이 크기가 비슷할 경우 pretraining에 사용하는 corpus 구성과 noise를 만드는 방법이 KLUE-TC benchmark에서 성능 개선에 큰 영향을 미치지 않고 오히려 성능 저하의 원인이 될 수 있음을 발견하였다.

이러한 실증적인 결과들을 통해, 우리는 실제 사람처럼 추론하는 한국어 뉴스 헤드라인 토픽 분류 모델을 만들기 위해서는 기존의 NSMC task에서 주로 사용하였던 pretraining 방법들, 즉, 한국어 모델의 pretraining에 사용되는 noise generation 방식과 학습에 사용하는 corpus 구성 방법등에 대해 깊은 고찰이 반드시 이루어져야한다고 생각했다.

향후 연구로는 KLUE-TC benchmark 성능 검증에 사용된 KLUE-BERT와 KLUE-RoBERTa에 대해서도 비슷한 표준 benchmark 성능 분석을 통해 RoBERTa에서 NSP task가 성능 개선에 큰 영향이 없다는 것을 발견한 것처럼 한국어 large-scale language model에서도 이러한 심층적인 분석을 진행하려고 한다.

참고문헌

[1] K. Cho, “Natural language understanding with distributed representation,” *CoRR*, Vol. abs/1511.07916, 2015. [Online]. Available: http://arxiv.org/abs/1511.07916

[2] A. Ahmadvand, H. Sahijwani, J. I. Choi, and E. Agichtein, “Concet: Entity-aware topic classification for open-domain conversational agents,” *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’19, p. 1371–1380, 2019. [Online]. Available: https://doi.org/10.1145/3357384.3358048

[3] J. Li, A. Sun, J. Han, and C. Li, “A survey on deep learning for named entity recognition,” *CoRR*, Vol. abs/1812.09449, 2018. [Online]. Available: http://arxiv.org/abs/1812.09449

[4] M. Namazifar, A. Papangelis, G. Tür, and D. Z. Hakkani-Tür, “Language model is all you need: Natural language

understanding as question answering," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7803–7807, 2021.

[5] W. Xu, B. Haider, and S. Mansour, "End-to-end slot alignment and recognition for cross-lingual NLU," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., pp. 5052–5063, 2020. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.410

[6] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., Vol. 119, pp. 11 328–11 339, 13–18 Jul 2020. [Online]. Available: https://proceedings.mlr.press/v119/zhang20ae.html

[7] C. Si, Z. Yang, Y. Cui, W. Ma, T. Liu, and S. Wang, "Benchmarking robustness of machine reading comprehension models," *Findings of ACL*, 2021.

[8] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., Vol. 28, 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

[9] S. Park, J. Moon, S. Kim, W. I. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, S. Kim, K. Lim, J. Lee, K. Park, J. Shin, S. Kim, L. Park, A. Oh, J. Ha, and K. Cho, "Klue: Korean language understanding evaluation," 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., pp. 4171–4186, 2019. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019, cite arxiv:1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692

[13] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2018. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[15] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[17] B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo, H. Lee, M. Jeong,

S. Lee, M. Kim, S. H. Ko, S. Kim, T. Park, J. Kim, S. Kang, N.-H. Ryu, K. M. Yoo, M. Chang, S. Suh, S. In, J. Park, K. Kim, H. Kim, J. Jeong, Y. G. Yeo, D. hyun Ham, D. Park, M. Y. Lee, J. Kang, I. Kang, J.-W. Ha, W. Park, and N. Sung, "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers," *ArXiv*, Vol. abs/2109.04650, 2021.

[18] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

[19] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.