KoBART와 GSG를 결합한 지능형 한국어 문장 요약 기법

심현솔, 박현빈, 박지영, 신재원, 김영종
숭실대학교 소프트웨어학부
soline013@gmail.com, phyeonbin01@gmail.com, jee_young___@naver.com,
t3qquq@gmail.com, youngjong@ssu.ac.kr

Intelligent Korean Sentence Summarization Technique Combining KoBART and GSG

Hyeonsol Sim, Hyeonbin Park, Jeeyoung Park, Jaewon Sin, Youngjong Kim
School of Software, Soongsil University

요 약

본 논문에서는 한국어 데이터와 모델링, 추가 평가 지표를 통해 Text Summarization 분야에서 한국어로 좋은 성능을 내기 위한 방식을 제안한다. KoBART의 크기를 키우고 PEGASUS의 GSG를 사용하는 KoBART-GSG 모델을 제안한다. 이때 ASR 모델을 사용하여 한국어 데이터를 구축하고 추가 학습을 진행한다. 또한, 생성된 요약문과 원문에서 Attention 기법으로 키워드와 핵심 문장을 추출하여 지능형 텍스트를 구성하는 새로운 방식을 제안한다. ASR Open API와 제안한 방식을 사용하여 오디오 파일을 텍스트로 변환하고 요약하는 강의나 회의 등 학계와 산업에서 사용할 수 있는 서비스를 제공한다.

1. 서론

텍스트 요약과 지능형 텍스트 구성은 현대 사회에서 정보 처리의 효율성과 정확성을 높이는 데 중요한 연구 주제이다. 최근 ChatGPT가 텍스트 요약을 비롯한 많은 자연어 처리 분야에 활용되고 있지만, 텍스트 요약에 맞게 사전학습된 모델이 부족하다는 한계를 가지고 있다. 또한, 다양한 도메인에 대한 체계적인 평가 방법이 부족한 상태이다. 본 논문은 이러한 문제점을 해결하기 위해 한국어 데이터와 모델링, 추가 평가 지표를 통한 방식을 제시한다. 기존 KoBART의 크기를 키우고, PEGASUS의 GSG를 적용한 KoBART-GSG를 제안한다. 또한, ASR 모델을 사용하여 추가 한국어 데이터를 구축하고, Attention 기법으로 키워드와 핵심 문장을 추출한다. 제안한 방식으로 향후 다양한 응용 분야에서 더 효율적이고 정확한 텍스트 요약이 가능해질 것으로 예상된다.

2. 관련 연구

Text Summarization은 중요한 정보와 의미를 유지하면서 긴 텍스트를 짧은 텍스트로 요약하는 자연어처리 활용 분야이다. Transformer(Vaswani et al., 2017)의 등장 이후, 요약 분야에도 Transformer가 적용되기 시작하였다. BertSum(Liu et al., 2019)은 사전학습 된 BERT에 Transformer Layer를 추가하여 요약 분야에 적용하는 방식을 제안하는 논문이다. BART(Lewis et al., 2020)는 BERT와 GPT를 seq2seq 구조로 사용하는 방식을 제안한다. BART는 Denoising Auto-encoder로, 임의의 노이즈를 학습 데이터에 적용한다. BART에서 요약 분야의 성능이 많이 향상되었고, CNN/DailyMail에서 SOTA를 달성하였다. HAT-BART(Rohde et al., 2021)는 HAT(Hierarchical Attention Transformer)을 BART에 적용한다. BART-LS(Xiong et al., 2022)는 Block Attention and Pooling Layer, Long Sequence from C4 Corpus, T5 Denoising과 같은 3가지 기법을 BART에 적용할 것을 제안한다. BART+R3F(Aghajanyan et al., 2020)는 BART에 R3F(Robust Representations through Regularized Finetuning) 기법을 적용한다. PEGASUS(Zhang et al., 2020)는 Transformer 구조로, MLM(Masked Language Model)과 다르게 문장 자체를 Masking

하는 GSG(Gap-Sentence-Generation)을 사용한다. PEGASUS는 요약 중에서도 Abstractive Summarization에 중점을 두고 연구되었다. 적은 비용으로 높은 성능을 나타내어 12개 데이터셋 중 6개에서 SOTA를 달성하였다. PEGASUS-X(Phang et al., 2022)는 Long Sequence에 좋은 성능을 내기 위한 방식을 제안한다. 최대 16K Token에 대해 적용할 수 있다. PEGASUS+DotProd(Kedia et al., 2021)는 PEGASUS에 Meta Dot Product를 사용하여 GigaWord Dataset에서 SOTA를 보인다.

3. 모델

3.1 KoBART

KoBART는 SKT에서 공개한 한국어 BART 모델이다. 40GB 이상의 한국어 데이터에 대해 사전학습 하였다. 학습 데이터는 한국어 위키 백과가 5M, 모두의 말뭉치v1.0, 청와대 국민청원 등의 데이터가 0.27B이다. 대화에 자주 사용되는 이모티콘, 이모지를 추가하여 해당 토큰의 인식 능력을 올렸다. 모델은 KoBART-base만 존재하며, Parameter 수는 124M이다. Encoder는 6개의 레이어를 사용하였고, Head는 16개, FFN 레이어는 3072, Hidden 레이어는 768이다. Decoder 환경은 Encoder와 동일하다. KoBART-base는 Classification, Regression 분야에서 NSMC 90.24, KorSTS 81.66, Question Pair 94.34를 달성하였다. KoBART-base의 공개 이후, 다양한 모델이 많은 연구자에 의해 제안되었다. 그중 KoBART-summarization은 Dacon 한국어 문서 생성요약 AI 경진대회의 학습 데이터를 활용하여 Fine Tuning 한 모델이다. KoBART-summarization의 상세한 아키텍처와 Performance는 공개된 자료가 존재하지 않는다.

3.2 KoBART-GSG

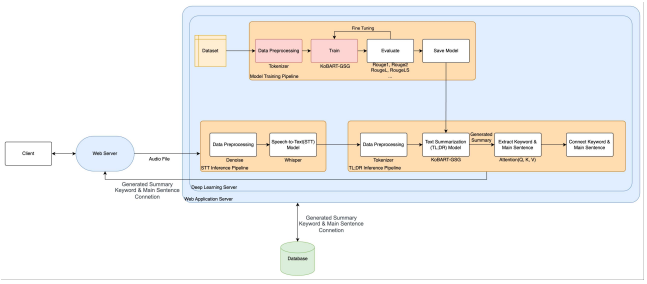

그림 1 전체 아키텍처

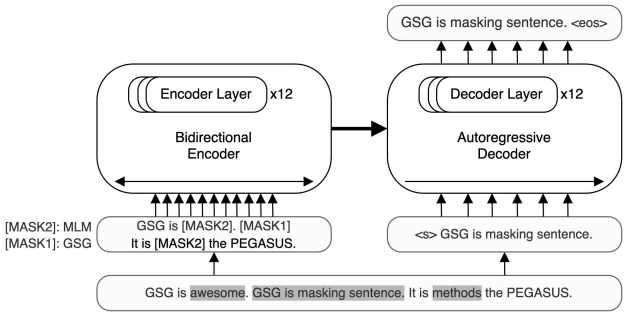

그림 2 KoBART-GSG 아키텍처

KoBART에 적용할 수 있는 다양한 아키텍처 중 몇 가지 방식을 사용하여 구현한다. 추가 데이터를 통해 Fine Tuning을 진행하고 파라미터를 튜닝한다. 이때 ASR 모델로 추가 데이터를 구축한다. KoBART의 크기를 늘린 KoBART-Large를 제안한다. Parameter 수는 BART-Large에 맞게 406M으로 설정하였다. 6개 레이어에서 12개 레이어로 크기를 늘리고, Head는 16개, FFN Dimension은 4096으로 BART-Large에서 공개된 수치는 아니나 PEGASUS를 참고하였다. Hidden Dimension은 1024이다. BART와 PEGASUS의 사이즈는 대부분 유사하다. PEGASUS의 GSG(Gap-Sentence-Generation)를 사용하여 문장 자체에 마스킹 혹은 노이즈를 적용한다. 기존 MLM(Masked Language Model) 방식도 병행하여 사용하는데, MASK2는 MLM, MASK1은 GSG를 사용한다. 생성된 요약문과 원문을 사용하여 Attention 기법으로 키워드와 핵심 문장을 추출하는 방식을 추가로 제안한다.

3.3 데이터셋

본 논문에서는 추가 데이터를 ASR 모델로 구축하는 방식을 제안한다. 이때 사용되는 모델인 Whisper는 OpenAI에서 발표한 ASR 모델이다. Transformer 구조를 사용하였고, 학습 데이터로 680,000 시간을 사용하였다. 그중 한국어 학습 데이터는 8,000 시간으로 영어를 제외하고 7번째로 많은 데이터를 사용하였다. 추가로 AIHub, Dacon에서 제공하는 요약 학습 데이터를 사용한다.

3.4 전처리

추가 데이터를 ASR 모델로 구축하는 방식은 데이터의 정확성을 보장하지 않는다. 따라서 정확성을 보장하기 위해 띄어쓰기 혹은 맞춤법을 교정하는 전처리 과정을 추가한다. 또한 ASR 모델 이전에 Denoise 모델이나

필터를 사용하여 음성 인식률을 최대한으로 높이는 방식을 사용한다.

4. 결론

KoBART 모델의 크기를 늘리고, PEGASUS의 GSG 기법을 적용한 KoBART-GSG를 제안한다. Denoising Auto-encoder와 문장 단위로 마스킹하는 GSG의 결합이 텍스트 요약에서 유망한 모델이 되길 기대한다.

참고문헌

[1] Mihalcea, R., & Tarau, P. “Textrank: Bringing order into text.” In *Proceedings of the 2004 conference on empirical methods in natural language processing,* pp. 404-411. 2004.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. “Attention is all you need.” *Advances in neural information processing systems*, *30.* 2017.

[3] Liu, Y., & Lapata, M. “Text Summarization with Pretrained Encoders.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP,* pp. 3730-3740. 2019.

[4] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* pp. 7871-7880. 2020.

[5] Rohde, T., Wu, X., & Liu, Y. “Hierarchical learning for generation with long source sequences.” *arXiv preprint arXiv:2104.07545*. 2021.

[6] Xiong, W., Gupta, A., Toshniwal, S., Mehdad, Y., & Yih, W. T. “Adapting Pretrained Text-to-Text Models for Long Text Sequences.” *arXiv preprint arXiv:2209.10052*. 2022.

[7] Aghajanyan, A., Shrivastava, A., Gupta, A., Goyal, N., Zettlemoyer, L., & Gupta, S. “Better Fine-Tuning by Reducing Representational Collapse.” In *International Conference on Learning Representations. 20202460.*

[8] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.” In *International Conference on Machine Learning,* PMLR, pp. 11328-11339. 20202460.

[9] Phang, J., Zhao, Y., & Liu, P. J. “Investigating Efficiently Extending Transformers for Long Input Summarization.” *arXiv preprint arXiv:2208.04347*. 2022.

[10] Kedia, A., Chinthakindi, S. C., & Ryu, W. “Beyond Reptile: Meta-Learned Dot-Product Maximization between Gradients for Improved Single-Task Regularization.” In Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 407-420. 2021.

[11] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. “Robust speech recognition via large-scale weak supervision.” *arXiv preprint arXiv:2212.04356*. 2022.