언어 번역 모델을 통한 한국어 지시 학습 데이터 세트 구축

임영서*, 추현창°*, 김산, 장진예, 정민영, 신사임
한국전자기술연구원
{ys_lim, cngusckd, kimsan0622, jinyea.jang, minyoung.jung, sishin}@keti.re.kr

Korean Instruction Tuning Dataset

Yeongseo Lim*, HyeonChang Chu°*, San Kim, Jin Yea Jang, Minyoung Jung, Saim Shin
Korea Electronics Technology Institute

요 약

최근 지시 학습을 통해 미세 조정한 자연어 처리 모델들이 큰 성능 향상을 보이고 있다. 하지만 한국어로 학습된 자연어 처리 모델에 대해 지시 학습을 진행할 수 있는 데이터 세트는 공개되어 있지 않아 관련 연구에 큰 어려움을 겪고 있다. 본 논문에서는 T5 기반 한국어 자연어 처리 모델인 Long KE-T5로 영어 데이터 세트를 번역하여 한국어 지시 학습 데이터 세트를 구축한다. 또한 구축한 데이터 세트로 한국어로 사전 학습된 Long KE-T5 모델을 미세 조정한 후 성능을 확인한다.

주제어: 지시 학습, 한국어 데이터 세트, 영-한 번역

1. 서론

자연어 처리 모델은 최근 몇 년 동안 상당한 발전을 이루고 있다. T5 [1]와 GPT-3 [2]같이 사전 학습된 거대 언어 모델은 많은 양의 데이터를 기반으로 훈련되어 다양한 작업에 대해 우수한 성능을 보이고 있다. 그러나 이러한 거대 언어 모델은 단순히 방대한 양의 데이터를 외우고 그 결과를 재현하는 방식으로 작동되기 때문에, 문맥에 맞는 대답을 얻어내기 쉽지 않다. 위의 문제점을 해결하기 위해 지시 학습을 진행하여 거대 모델을 미세 조정하고 있다.

지시 학습(Instruction tuning) [3]이란 언어 모델이 사용자의 질의와 명령에 적합한 대답을 생성도록 하기 위하여 데이터를 템플릿을 이용하여 명령 프롬프트-대답 쌍으로 구성한 후 모델을 미세 조정하는 방식이다. 최근 지시 학습을 사용하여 미세 조정한 자연어 처리 모델은 상당한 성능 향상을 보인다. 그러나 이러한 지시 학습을 위한 데이터 세트는 영어 데이터를 기반으로 구축되어 있기 때문에 한국어 자연어 처리 모델에 적용하기가 어려운 상황이다.

본 논문에서는 이러한 제한을 극복하기 위해 영어 데이터 세트를 번역 후 템플릿과 결합하여 한국어지시 학습 데이터 세트를 구축한다. 번역에는 T5 기반의 한국어 자연어 처리 모델인 Long KE-T5 [4][1]를 사용하였다. 또한, 구축된 데이터를 사용하여 한국어로 사전 학습된 Long KE-T5 모델을 미세 조정하고 그 성능을 확인했다.

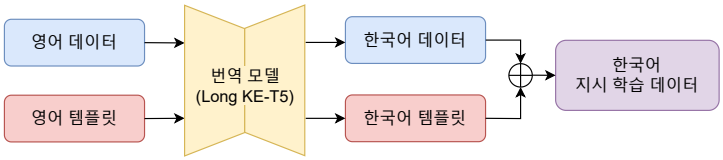

그림 1. 한국어 지시 학습 데이터 구축 과정

2. 최근 연구

2.1 대형 언어 모델

2.1.1 T5

T5는 Transformer [5] 기반의 인코더-디코더 모델로 다양한 자연어 처리 작업을 진행하기 위한 하나의 통일된 프레임워크를 제공한다. 이 덕분에 모델의 일반화 능력이 향상되어서 다양한 자연어 처리 작업에 활용할 수 있다. 특히 질문 응답, 요약, 대화 생성 등 자연어 이해뿐만 아니라 생성 작업에도 좋은 성능을 보이고 있다.

2.1.2 KE-T5&Long KE-T5

KE-T5 [6]는 T5 모델을 한국어와 영어 말뭉치로 사전 학습한 모델로 대부분의 한국어 자연어 처리 작업을 수행할 수 있는 모델이다. 해당 모델은 많은 양의 한국어와 영어 어휘를 포함하는 대규모 언어 모델로 다양한 한국어 처리 작업에서 높은 성능을 보인다. 또한 입력길이를 늘린 Long T5[7]를 기반으로 한국어와 영어 말뭉치를 활용한 Long KE-T5도 공개되었으며 KE-T5보다 더 긴 문맥을 입력으로 받을 수 있는 특징이 있다.

*공동저자(These authors contributed equally)
[1]https://github.com/AIRC-KETI/long-ke-t5

처리 분야	데이터 세트
텍스트 생성	WikiBio [9]
요약	XSum [10], SAMSum [11]
다지선다 질의응답	RACE [12]
감성 분석	App_reviews [13], IMDb [14]
주제 분류	AG_News [15], TREC [16, 17]

표 1. 구축에 활용된 데이터 세트

2.2 지시 학습

2.2.1 FLAN

FLAN [3]은 거대 언어 모델을 지시 학습으로 미세 조정하여 제로샷(Zero shot) 성능을 높인 모델이다. 기존의 언어 모델에서는 모델이 출력해야 할 작업 목표를 데이터의 앞쪽에 토큰처럼 붙여서 입력으로 넣어주거나 태스크 별로 다운스트림 모델을 따로 학습하였다. 하지만 지시 학습은 대형 언어 모델들이 사용자의 요구에 맞는 대답을 하도록 튜닝하는 것이 목적이기 때문에, 데이터 세트를 명령형 프롬프트와 대답 쌍으로 가공하고 이를 활용하여 모델을 학습시킨다.

2.2.2 T0

T0 [8]는 지시 학습으로 T5 모델을 미세 조정하여 제로 샷 성능을 높인 모델이다. 해당 모델에서 사용한 템플릿은 FLAN의 템플릿보다 더 다양한 상황을 제시한다. 지시 학습 데이터 세트를 만들기 위해서는 원본 데이터와 데이터를 가공하기 위한 틀인 템플릿이 필요하다. 본 논문에서는 FLAN과 T0에서 사용한 템플릿을 번역하여 한국어 지시 학습 데이터 세트를 만드는 데에 사용했다.

3. 한국어 지시 학습 데이터 세트

3.1 구축에 활용된 데이터 세트

데이터 구축에 활용할 데이터 세트는 총 5개의 자연어 처리 분야에서 각각 1~2개의 데이터 세트를 선정하였으며 각 자연어 처리 분야와 선정된 데이터 세트는 표 1와 같다. WikiBio[9]는 영어 위키피디아에서 표로 구성되어있는 인물에 대한 정보를 문어체로 변환하는 태스크다. XSum[10]과 SAMSum[11]은 각각 문서와 대화를 요약하는 태스크이며, RACE[12]는 다지선다로 구성된 중국의 중고등학교 영어시험 데이터이며, App reviews[13]와 IMDb[14] 데이터 세트는 각각 안드로이드 앱과 영화의 리뷰 및 평점 데이터 세트이다. 마지막으로 AG News 데이터 세트와 TREC 데이터 세트는 각각 뉴스의 주제와 질의의 종류를 구분하는 태스크로 구성되어있다.

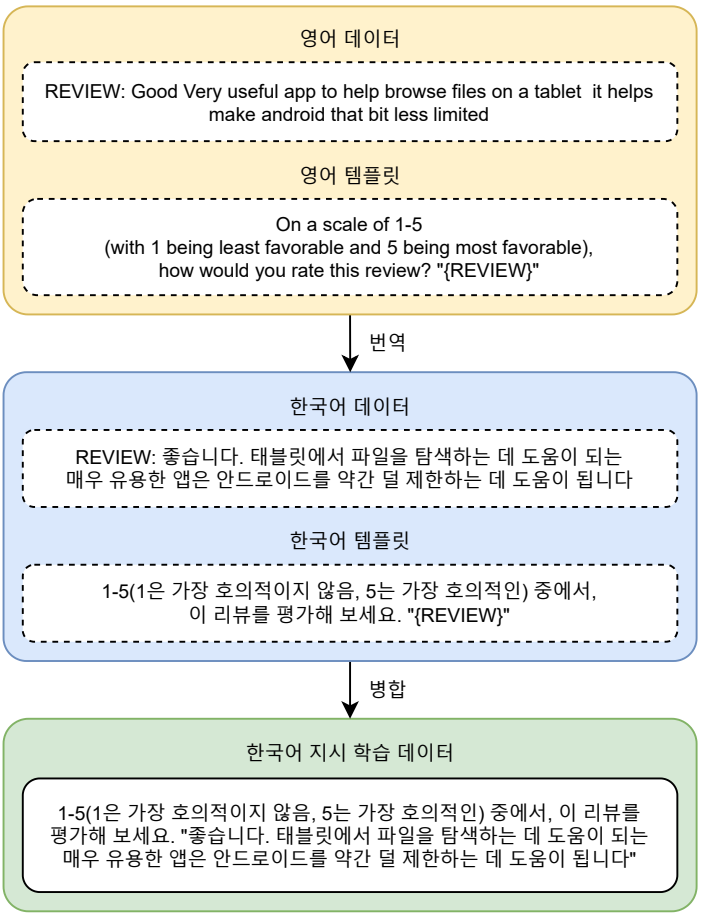

그림 2. 감성 분석 태스크 샘플의 한국어 지시학습 데이터 구축 예시

3.2 데이터 세트 구축 방법

허깅페이스[2]에 공개된 영어 데이터 세트 중 감성 분석, 요약 등 각 자연어 처리 분야에 맞는 데이터 세트와 FLAN과 T0에 공개된 영어 템플릿을 사용했다. 그림 1은 영어 자연어 처리 데이터와 템플릿을 이용하여 한국어 지시 학습 데이터 세트를 구축하는 과정을 보여준다. 먼저 영어 데이터 세트의 텍스트를 문장 단위로 분할하고, 나누어진 문장들을 각각 Long KE-T5 번역 모델[3]을 이용해 한국어로 번역 후 다시 하나의 텍스트로 연결하여 한국어 데이터 세트를 구축했다.

영어 템플릿은 처리 분야에 맞지않는 템플릿은 제거한 뒤 Long KE-T5 모델을 이용해 번역한 후 어순이나 어법이 어색한 경우 원래 영어 템플릿이 가지고 있는 의미를 유지하도록 수정했다. 이렇게 수정된 한국어 템플릿을 구어체나 존칭 어체와 같은 어체로 변경하여 한국어 템플릿을 증강했다. 번역된 한국어 데이터 세트와 증강된 한국어 템플릿을 병합하여 한국어 지시 학습 데이터 세트를 구축했다. 그림 2는 감성분석

[2] https://huggingface.co/
[3] https://huggingface.co/KETI-AIR-Downstream/long-ke-t5-base-translation-aihub-bidirection

텍스트 생성	
입력:	이 집합에 대한 문장을 생성하시오: ['물리학, ph. d.', '1965년 10월 5일', '뮌헨, 독일', '작가님.', '베를린.', 'stefan klein portrait.jpg', '뮌헨 대학', '스테판 클린', '오스트리안.', 'www.stefanklein.info', '스테판 클린']
출력:	스테판 클레인(1965년 10월 5일생)는 물리학자, 작가, 수필가로 '행복의 과학'과 '시간: 사용자 가이드'로 가장 잘 알려져 있다.
요약	
입력:	이 대화를 요약하면? **엘라**: 조금 이따가 올게 **김**: 알았어 **홀리**: 버스 놓쳤어? **엘라**: 언제나 처럼
출력:	엘라는 버스를 놓쳤고 조금 늦을 거예요.
질의응답	
입력:	나는 시험을 보고 있는데 기사 다음에 나오는 질문을 맞춰야 해. 기사: 사람들은 음식에 대해 이상한 생각을 가지고 있어요.예를 들어, 토마토는 매우 맛있는 채소의 일종입니다. 여러 가지 방법으로 준비할 수 있는 유용한 식물 중 하나입니다.풍부한 영양과 비타민이 함유되어 있습니다. 그러나 18세기에 미국인들은 토마토를 먹지 않았습니다.토마토 식물이 너무 예뻐서 정원에서 키웠어요.그러나 그들은 야채가 독이 있다고 생각했습니다. ... 대통령은 그의 요리에 토마토 크림 수프를 만드는 방법을 가르쳤다. 손님들은 수프가 정말 맛있다고 생각했어요. 질문: 지문을 읽고 나면 다음 중 어느 것이 사실이라고 생각하는가? 선택지: A: 미국인들은 토마토를 심기 시작한 후 결코 토마토를 먹지 않았다. B: 미국인들은 19세기 이전에 토마토를 먹지 않았어요. C: 지금도 미국인들은 토마토를 먹지 않아요. D: 18세기에 미국인들은 토마토를 많이 먹었습니다.
출력:	B
감성 분석	
입력:	놀라운 윌리엄스 씨에 대해서는 놀라운 것이 없습니다.이 영화의 문제 중 일부는 주연 배우 멜빈 더글라스이다. 그는 형편없는 배우였고 게으른 배우였다. ... 당신은 멜빈 더글라스를 드랙으로 보고 싶지 않을 것입니다!그는 6피트가 훨씬 넘었고, 그 짜증나는 콧수염도 면도하지 않았어요. ...그것이 바로 '놀라운 미스터 윌리암스'의 문제, 즉 모든 것이 너무 명백하다는 것이다.이 영화를 10점 만점에 2점으로 평가할게요. 영화에 대해 표현된 감정은
출력:	부정적인 감정입니다
주제 분류	
입력:	영웅들은 따뜻한 환영을 받는다. BBC 스포츠의 댄 워런은 올림픽에서 홈팀을 환영하면서 팬들과 합류한다. 다음의 신문 섹션 중 이 기사가 나타날 가능성이 높은 것은? 월드 뉴스, 스포츠, 비즈니스, 과학 기술?
출력:	스포츠

표 2. 처리 분야별 구축된 데이터 세트 예시

데이터 세트인 app reviews 데이터 샘플을 한국어 데이터 세트로 변환하는 과정을 보여준다. 표 2은 각 처리 분야별로 언급한 방법으로 구축된 데이터 샘플을 보여준다.

4. 실험

4.1 실험 방법

각각의 처리 분야에서 6,000개 씩 추출한다. 한 처리 분야의 데이터 세트가 2개 이상인 경우는 데이터 세트의 크기에 비례하여 추출한다. 각각의 처리 분야 마다 69 : 1 : 30 비율로 학습, 검증, 테스트 세트로 분할한다. 그 후 번역된 FLAN 템플릿과 T0 템플릿에서 총 합 10개가 되도록 템플릿을 추출한 뒤 각 데이터마다 한국어 템플릿에 적용시킨다. 이렇게 구성된 학습 데이터 세트를 이용해 Long KE-T5 모델을 미세 조정하고 테스트 세트를 이용해 BLEU 점수[18]와 ROUGE 점수[19]로 모델을 평가한다.

4.2 실험 결과

통합한 데이터 세트에서 Long KE-T5의 실험 결과는 표 3과 같다. 처리 분야별 실험 결과는 표 4와 같다.

모델	BLEU-1	ROUGE-1	ROUGE-2
Long KE-T5	56.80	52.90	28.09

표 3. 통합된 데이터 세트 실험 결과

처리 분야	BLEU-1	ROUGE-1	ROUGE-2
텍스트 생성	56.66	39.46	26.00
요약	27.44	25.90	7.54
질의 응답	42.78	42.09	21.48
감성 분석	66.11	66.11	12.86
주제 분류	91.03	90.92	72.59

표 4. 처리 분야별 데이터 세트 실험 결과

4.3 실험 결과 분석

4.3.1 정량적 평가 결과

표 3의 실험결과에서 ROUGE-1의 점수가 ROUGE-2의 점수보다 약 1.8배 높음을 확인할 수 있다. 이는 처리 분야 중 텍스트 생성과 요약을 제외한 나머지 처리 분야의 답이 단답형인 경우가 많아 ROUGE-2 점수가 낮게 측정되었기 때문으로 보인다. 실제로 표 4를 보면 주로 단답형 답이 많은 감성 분석은 ROUGE-1 점수가 ROUGE-2 점수보다 훨씬 높은 것을 확인할 수 있다.

요약의 경우 다른 처리 분야의 비해 정답을 구성하는 단어 수가 많으며 이로 인해 점수가 낮다. 그리고 주제 분류의 경우에는 정답으로 선택할 수 있는 선택지가 입력으로 주어지기 때문에 점수가 높다.

4.3.2 정성적 평가 결과

실제로 모델이 예측한 한 데이터의 예시를 보면 입력으로 "***놀라운*** *윌리엄스 씨에 대해서는* ***놀라운*** *것이 없습니다.이 영화의 문제 중 일부는 주연 배우 멜빈 더글라스이다.그는 형편없는 배우였고 게으른 배우였다. . . .그것이 바로 '****놀라운*** *미스터 윌리암스'의 문제, . . .이 영화를 10점 만점에 2점으로 평가할게요. 영화에 대해 표현된 감정은*" 이라는 문장이 들어왔을 때 출력으로 "긍정적인 감정입니다"를 예측했다. 이는 부정적인 감정을 가진 리뷰임에도 긍정적인 리뷰에 많이 들어가 있는 "놀라운" 이라는 단어가 많이 들어가 있었기 때문에 잘못 예측한 것으로 보인다.

5. 결론

본 논문에서는 새로운 한국어 지시 학습 데이터 세트을 구축하고 구축된 데이터 세트의 성능을 실험을 통하여 보였다. 학습된 모델은 다양한 자연어 처리 태스크에서 준수한 성능을 보였으나, 질의 응답과 요약에서는 상대적으로 낮은 성능을 보였다. 질의 응답과 요약에서의 성능을 더 높이기 위하여 향후 연구로 자연어 처리 분야를 확장하고 처리 분야마다 데이터 세트를 추가함으로서 거대 한국어 자연어 모델의 지시 학습을 위한 데이터 세트를 확보하고 거대 한국어 자연어 처리 모델의 제로 샷 성능을 평가하고자 한다.

감사의 글

이 논문은 2023년도 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원(No. 2022-0-00320)의 지원을 받아 수행된 연구임

참고문헌

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, Vol. 21(140), pp. 1–67, 2020.

[2] 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," *Conference on Neural Information Processing Systems*, 2020.

[3] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *International Conference on Learning Representations*, 2022.

[4] K. AIRC, "Long-ke-t5: Long korean english t5," May 2023. [Online]. Available: https://github.com/AIRC-KETI/long-ke-t5

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," *Conference on Neural Information Processing Systems*, 2017.

[6] S. Kim, J. Y. Jang, M. Jung, and S. Shin, "A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems," *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 352–365, Nov. 2021. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.33

[7] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang, "LongT5: Efficient text-to-text transformer for long sequences," *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 724–736, Jul. 2022. [Online]. Available: https://aclanthology.org/2022.findings-naacl.55

[8] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. T.-J. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. A. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf, and A. M. Rush, "Multitask prompted training enables zero-shot task generalization," *International Conference on Learning Representations*, 2022.

[9] R. Lebret, D. Grangier, and M. Auli, "Generating text from structured data with application to the biography domain," *CoRR*, Vol. abs/1603.07771, 2016. [Online]. Available: http://arxiv.org/abs/1603.07771

[10] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," *ArXiv*, Vol. abs/1808.08745, 2018.

[11] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization," *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Nov. 2019. [Online]. Available: https://www.aclweb.org/anthology/D19-5409

[12] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "RACE: Large-scale ReAding comprehension dataset from examinations," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Sep. 2017. [Online]. Available: https://aclanthology.org/D17-1082

[13] "Software applications user reviews," 2017.

[14] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, June 2011. [Online]. Available: http://www.aclweb.org/anthology/P11-1015

[15] X. Zhang, J. J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *NIPS*, 2015.

[16] X. Li and D. Roth, "Learning question classifiers," *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. [Online]. Available: https://www.aclweb.org/anthology/C02-1150

[17] E. Hovy, L. Gerber, U. Hermjakob, C.-Y. Lin, and D. Ravichandran, "Toward semantics-based answer pinpointing," *Proceedings of the First International Conference on Human Language Technology Research*, 2001. [Online]. Available: https://www.aclweb.org/anthology/H01-1069

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[19] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, pp. 74–81, Jul. 2004. [Online]. Available: https://aclanthology.org/W04-1013