The 6th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

A Dataset for Korean Graph-to-Text Generation

Dahyun Jung[1,5], Seungyoon Lee[2,5], SeungJun Lee[3], Jaehyung Seo[3], Sugyeong Eo[3], Chanjun Park[4], Titi[5*]

[1] Hanguk University of Foreign Studies
[2] Chung-ang University
[3] Korea University, Computer Science
[4] Upstage
[5] Human-inspired AI Research
{dhaabb14, dltmddbs100}@gmail.com
{dzzy6505, seojae777, djtnrud, titi}@korea.ac.kr
bcj1210@naver.com

Abstract. Recently, research on generating text based on knowledge graphs has been importantly dealt with. However, datasets for these studies are insufficient, and there are many limitations in building datasets. Therefore, this paper proposes a methodology for automatically building data based on published knowledge graphs. Based on this, a sentence generation experiment was conducted using a large-capacity language model including Korean such as KoBART and mBART.

Keywords: Graph to Text, Knowledge Graph, Natural Language Processing

1 Introduction

Graph-to-text generation is a task of generating text by receiving a graph as an input. Recently, research on these tasks has been actively conducted [1], and there is a limitation in that it takes a lot of time and money to build a dataset for the task. Therefore, this paper presents a plan to automatically build a Korean-based graph-to-text dataset. The proposed Korean graph-to-text dataset is a dataset based on DBpedia, an open knowledge graph, consisting of triplets representing the relationship between one entity and another and text pairs describing the entity. Based on the established dataset, experiments were conducted using KoBART[1], a Korean pre-learning model, and mBART [2], a multilingual language model.

2 Related Work

There are two main methods of graph-to-text generation: the first is to utilize graph neural networks or graph transformers [3]. The second method used in this paper is to linearize the graph and use it as a direct input [4].

* Corresponding Author
[1] https://github.com/SKT-AI/KoBART

Graph-to-text datasets typically have WebNLG [5]. WebNLG is data based on RDF triple composed of various domains. In addition, text describing these triplets is structured together, making it a suitable dataset for performing graph-to-text tasks. As such, graph-to-text data in English exists, but there is no dataset in Korean.

3 Methods

To automate the construction of graph-to-text datasets, we utilize Korean DBpedia data, a public knowledge graph. Text was extracted from DBpedia data, and a knowledge graph was constructed by selecting a triple related to the sentence extracted from the triple existing in DBpedia. The graph-to-text dataset constructed in this way consists of a knowledge graph describing an entity and a sentence associated with the graph.

Table 1. Dataset experimental results.

Model	BLEU-3	BLEU-4	Morpheme BLEU	METEOR	CHRF++	KoBERT Score
KoBART	14.69	11.38	24.1	38.68	**31.96**	82.31
mBART	**15.37**	**12.03**	**25.2**	**40.77**	**31.96**	**84.91**

4 Research result and Conclusion

Table 1 shows the experimental results for the graph-to-text dataset. The mBART model showed the best performance compared to KoBART. These experimental results show that language models can perform graph-to-text tasks well with graph-to-text datasets in this paper.

In this study, we recognized the necessity of the Korean graph-to-text generation task and suggested a method of generating natural language sentences based on Korean knowledge graphs. Automated graph-to-text dataset construction methods can be used to construct data at a low cost, which can serve as a solution to the dataset shortage problem faced.

5 Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

References

1. Bai, X., Chen, Y., & Zhang, Y. (2022). Graph Pre-training for AMR Parsing and Generation. arXiv preprint arXiv:2203.07836.
2. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726-742.
3. Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., & Hajishirzi, H. (2019). Text generation from knowledge graphs with graph transformers. arXiv preprint arXiv:1904.02342.
4. Ribeiro, L. F., Schmitt, M., Schütze, H., & Gurevych, I. (2020). Investigating pretrained language models for graph-to-text generation. arXiv preprint arXiv:2007.08426.
5. Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017, September). The WebNLG challenge: Generating text from RDF data. In Proceedings of the 10th International Conference on Natural Language Generation (pp. 124-133).