
Automatic Knowledge Augmentation for Generative Commonsense Reasoning

Jaehyung Seo
Korea University
seojae777@korea.ac.kr

Chanjun Park
Korea University
bcj1210@korea.ac.kr

Sugyeong Eo
Korea University
djtnrud@korea.ac.kr

Hyeonseok Moon
Korea University
glee889@korea.ac.kr

Heuiseok Lim[†]
Korea University
limhseok@korea.ac.kr

Abstract

Generative commonsense reasoning is the capability of a language model to generate a sentence with a given concept-set that is based on commonsense knowledge. However, generative language models still struggle to provide outputs, and the training set does not contain patterns that are sufficient for generative commonsense reasoning. In this paper, we propose a data-centric method that uses automatic knowledge augmentation to extend commonsense knowledge using a machine knowledge generator. This method can generate semi-golden sentences that improve the generative commonsense reasoning of a language model without architecture modifications. Furthermore, this approach is a model-agnostic method and does not require human effort for data construction.

1 Introduction

CommonGen [9] is a challenge in which the task is to generate a sentence that makes sense with an input concept-set by leveraging generative commonsense reasoning. However, it is difficult to derive much commonsense knowledge for the given concept-sets, such as compositional generalization and relational reasoning, from the 67,389 sentences in the CommonGen training set [5, 6]. Additionally, most reference sentences consist of superficial scene descriptions and do not represent any specific relational knowledge between the concepts [9].

To address these issues, we propose a data-centric approach that augments prior knowledge with an additional pre-training session. Our method uses a trained generator that produces semi-golden sentences as patterns to guide generative commonsense reasoning [8, 16]. As the generator, we utilize BART [7], which has the highest average percentage of input concepts that exist in the output sentences in the CommonGen task (97.35%). The generator learns to reconstruct sentences using concepts (e.g., verbs and nouns) extracted from sentences in Wikipedia articles, which have high coverage. Semi-golden sentences with concrete descriptions about the input concept-set or concept-pair set are used in an additional pre-training session to enhance the prior knowledge of the generative language model.

Automatic knowledge generation substantially reduces human effort and ensures the quality of machine-generated data using the results of performance improvements. It can also extend the relational information of given concept-sets, including commonsense knowledge for the intended input query. Moreover, the improvements in performance in auto-regressive models (*i.e.*, a decoder only) [12] and sequence-to-sequence language model [13] (*i.e.*, an encoder-decoder), which have

Table 1: Experimental results and statistics for each data augmentation used in pre-training. “Pair” and “Set” respectively indicate to the semi-golden sentences for concept-pairs and concept-sets in the CommonGen training set; “# sentences” and “# concept-sets” respectively refer to the number of semi-golden sentences and number of concepts with respect to the size of the concept-set.

	Results					Augmentation	
	ROUGE-L	BLEU4	METEOR	CIDEr	SPICE	# sentences	# concept-sets
GPT2	39.28	21.10	26.20	12.15	25.90	-	-
GPT2+Pair	49.70	23.10	26.90	12.62	27.20	59,125	2(59,125)
GPT2+Set	49.20	22.30	26.80	12.60	27.10	32,471	3(24,891), 4(4,206), 5(3,374)
GPT2+Pair+Set	49.50	23.50	27.20	12.95	27.60	91,596	2(59,125), 3(24,891), 4(4,206) 5(3,374)
T5	49.27	28.60	30.10	14.96	31.60	-	-
T5+Pair	56.90	33.70	33.00	17.72	33.80	59,125	2(59,125)
T5+Set	57.20	34.40	33.00	17.97	33.90	32,471	3(24,891), 4(4,206), 5(3,374)
T5+Pair+Set	56.80	33.40	32.90	17.76	33.80	91,596	2(59,125), 3(24,891), 4(4,206) 5(3,374)

different architectures, demonstrate the strong generalization and model-agnostic nature of this approach.

2 Data Construction Process and Results

In this section, we present the process of data construction and experiments using a generator.

First, we extracted 154,892 sentences from a Wikipedia data dump (2021-03-01) that match more than two input concepts via ElasticSearch [4] and BM25 [3, 14]. From the extracted sentences y , we defined concept-set C , which includes a verb or noun through part-of-speech (POS) tagging. **Second**, the generator was trained with the reconstructed data using $C = y$ and produced a semi-golden sentence y' for an input concept-set C' . **Third**, 59,125 concept-pairs and 32,471 unique concept-sets (of size 3 to 5) were extracted from the CommonGen training set and the generator used them as an input sequence to create a semi-golden sentence. **Fourth**, extracted concept-sets and generated semi-golden sentences were used as pre-training data in the form of $C' = y'$ for two generative language models, as presented in Table 1. We conducted a case study to demonstrate the impact of our method with GPT2 [12] and T5 [13]. To evaluate the performance, we used the n-gram overlapping metrics BLEU [11], ROUGE [10], and METEOR [2] as well as the concept matching metrics CIDEr [15] and SPICE [1].

As shown in Table 1, all data augmentation cases lead to performance improvements with respect to all evaluation metrics. The results also indicate that there is a difference in performance depending on the combination of input concepts and target models. Augmented concept-pairs include concrete relational knowledge between two concepts and concept-sets have commonsense knowledge for generating sentences by compositionality. GPT2 obtains the best performance improvement when learning both types of machine-generated knowledge through the augmented dataset. In contrast, T5 yields no significant performance difference depending on the form of the augmented dataset, although learning commonsense knowledge with compositionality is slightly more effective. This result shows that there is a variance in the degree of improvement in prior knowledge in a specific area when the model architecture and dataset used for pre-training each language model differ. This experimental result demonstrates that our data-centric approach generates better semi-golden sentences for use as an augmented dataset. It is also possible to produce a machine-generated dataset in the intended format to compensate for the insufficient prior knowledge of each model. Furthermore, this approach is a model-agnostic method and has a positive impact on both auto-regressive and sequence-to-sequence architectures.

3 Conclusion

We demonstrated how to use natural language generation for data-centric research to reduce the cost of leveraging information retrieval and building a human-annotated dataset. Automatic knowledge augmentation for generative commonsense reasoning is also not limited to a particular model or dataset, but can be applied in different forms as needed. In future work, this method will be highly valuable because it can be extended to other downstream tasks that require commonsense knowledge as supporting evidence.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and IITP grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). Heuseok Lim[†] is a corresponding author.

References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: semantic propositional image caption evaluation. In *European Conference on Computer Vision (14th: 2016)*, pages 382–398. Springer, Springer Nature.
- [2] Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- [3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.
- [4] Manda Sai Divya and Shiv Kumar Goyal. 2013. Elasticsearch: An advanced and quick search technique to handle voluminous data. *CompuSoft*, 2(6):171.
- [5] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.
- [6] Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2873–2882. PMLR.
- [7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- [8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- [9] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1823–1840.
- [10] Chin-Yew LIN. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain*, pages 74–81.
- [11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- [12] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- [13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

- [14] Krysta M Svore and Christopher JC Burges. 2009. A machine learning approach for improved bm25 retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1811–1814.
- [15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- [16] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. G-daug: Generative data augmentation for commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1008–1025.