Unified Denoising Pretraining and Finetuning for Data and Texts

Jiayi Xian[1*] **Dingcheng Li**[2] **Alexander Hanbo Li**[3]
Derek Liu[2] **Xing Fan**[2] **Chenlei Guo**[2] **Yang Liu**[2] **Yuqing Tang**[2,4*]
[1]University at Buffolo [2]Amazon Alexa AI [3]Amazon AWS AI [4]XPeng Motors X-Lab
jxian@buffalo.edu
{lidingch,hanboli,derecliu,fanxing,guochenl,yangliud}@amazon.com
tangyq@xiaopeng.com

Abstract

Text-to-Text (T2T) denoising-pretraining-finetuning (DPF) paradigms (e.g. BERT, BART, GPT) have achieved great success in a wide range of encoding and decoding tasks in NLP. However, little has been explored on data-to-data (D2D) and data-to-text (D2T) tasks using DPF paradigms. This work fills in the gap by investigating D2D and T2T denoising-pretraining for D2T tasks. D2D and T2T DPF paradigms can leverage large amount of unlabeled structural data (e.g. knowledge sub-graphs, annotated triples and other forms) and texts to train large capacity models. With the proposed T2T and D2D denoising pretraining, we improve the state-of-the-art performance of D2T tasks by 1.3 and 0.8 BLEU points for WebNLG and E2ENLG evaluation respectively.

1 Introduction

Text-to-Text (T2T) denoising-pretraining-finetuning (DPF) paradigms, e.g. BERT (Devlin et al., 2019), BART (Lewis et al., 2020), GPT (Radford et al., 2018), T5 (Raffel et al., 2020) and so on, have achieved great success pushing the state-of-the-art of NLP into a new level. On the other hand, Data-to-Text (D2T) – generating texts from structural data — a long lasting important task for various applications: e.g. dialog responses (Gu et al., 2021), news and weather reports (Clare et al., 2021), is not fully studied under DPF paradigms. Recently there are works on leveraging text pretraining for D2T, e.g. D2T with BERT (Lewis et al., 2020), D2T with BART (Lewis et al., 2020), KGPT (Chen et al., 2020). However, little has been explored regarding the possibility of leveraging pretraining on structural data (e.g. knowledge sub-graphs, annotated triples and etc.) without labels.

In this work, we investigate into the idea of (1) data-to-data (D2D) denosing pretraining and its relationship to (2) text-to-text (T2T) denoising pretraining with D2T as end tasks. To leverage DPF paradigms for structural data, we convert structural data triples (and the underlying sub-graphs) into a sequence of tokens with tags following KGPT (Chen et al., 2020). After structural flattening, the same transformer encoder and decoder can be applied for both texts and structural data. We test our proposed approach with WebNLG (RDF-to-text and text-to-RDF) (Shimorina and Gardent, 2018) and E2ENLG datasets (end-to-end (E2E), data-driven NLG) (Dusek et al., 2020). Firstly we found that new D2T SOTA can be achieved simply by finuning from a one-model-for-all multilingual translation model, specifically from an mBART50 (Tang et al., 2021) — a single translation model supporting 50 languages created by DPF paradigms. Secondly we found that with additional mixed D2D and T2T denoising pretraining, D2T performance can be further improved. Our contributions are of three-folds: (1) investigatigation of unified models for T2T, D2D and D2T tasks; and (2) extending DPF paradigms into D2D as well as joint D2D and T2T denoising pre-traning; (3) achieving new WebNLG and E2ENLG SOTA.

2 Methodology

Our approach is composed of three major ingredients: (1) flatten knowledge sub-graphs as tagged sequences of triples following KGPT (Chen et al., 2020); (2) unifying encoder-decoder transformer (Vaswani et al., 2017) for both texts and tagged triples; as a result (3) enabling T2T, D2T and D2T pretraining and finetuning using a mixed dataset of unlabelled structural data and texts as well as a supervised finetuning dataset.

2.1 Encoding and decoding structural data

We take structural data broadly to include any format of data that can be tagged into a sequence of tokens. This includes knowledge graphs, triples, ta-

*Work was done at Amazon Alexa AI.

ble rows and so on, by following the data flattening strategy employed in KGPT (Chen et al., 2020). It can be easily adapted to the graph representation of data using graph transformer (Yun et al., 2019). We will leave this as our future work.

2.2 Unifying Data-to-Data, Text-to-Text and Data-To-Text models

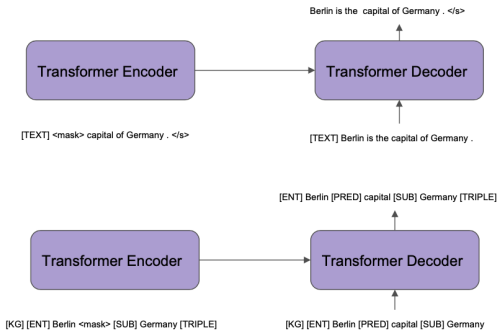

Figure 1: T2T and D2D denoising pretraining

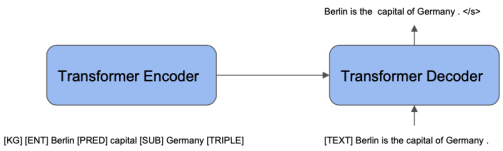

Figure 2: D2T finetuing

We unify Data-to-Data (D2D), Text-to-Text (T2T) and Data-to-Text (D2T) models into the same encoder-decoder architecture as shown in Figure 1 and Figure 2. Figure 1 depicts mixture of denoising D2D and T2T pretraining. During pretraining, we adapt the masking strategy used in mBART (Liu et al., 2020) with additional masking of the tags, the whole entity, relationship names with probabilities as hyper-parameters. Figure 2 depicts the finetuning strategy for a D2T end task.

3 Experiments

Different model initialization strategies are experimented using standard WebNLG and E2ENLG datasets as evaluation. Results are reported in Table 1. Among them, *mbart.rnd* is the same architecture as mBART50 from (Tang et al., 2021) but initiated randomly. *mbart.rnd.DTD* has the same architecture but it is initiated with mixed D2D and T2T denoising pretraining using unlabeled texts and triples extracted by KGPT (Chen et al., 2020). *mbart.rnd.MD2TF* is initiated with D2T supervised training using the heuristicly mined D2T dataset from KGPT. *mbart50.nn* is the mBART50 model from (Tang et al., 2021) [1] — a translantion model supporting 50 languages trained by DPF paradigms. *mbart50.nn.DTD* is the same mBART50 model but further denoisingly pretrained with a mixed unlabeled data and text dataset extracted by KGPT. We adapt the mBART implementation from (Tang et al., 2021) and augment its vocab with tagging symbols (e.g. [ENT], [SUB], and etc.) from KGPT. A few observations are found. Firstly we found that finetuning from the text-trained *mBART50.nn* improves the D2T SOTA by 1.1 and 0.7 BLEU points on WebNLG and E2ENLG respectively. This indicates that large-scale text DPF training can be transferred to D2T tasks well. Secondly we found that with additional mixed D2D and T2T denoising pretraining, D2T performance is further improved by 0.2 BLEU for WebNLG. Thirdly, compared to training D2T models from scratch, mixed D2D and T2T pretraining significantly improves D2T performance. Fourthly, text-only denoising pretraining achieves better performance than previous SOTA for both WebNLG and E2ENLG. As our unlabeled data and text dataset from KGPT are roughly about 700K sub-graphs and sentences respectively, which are much smaller than the mBART50 pretraining data, further study is needed to fully understand the potential of data denoising pretraining.

Model initialization	WebNLG	E2ENLG
KGPT (paper)	64.11	68.05
mbart.rnd	57.16	64.25
mbart.rnd.MD2TF	62.58	66.34
mbart.rnd.DTD	63.24	67.56
mbart50.nn	65.25	**68.85**
mbart50.nn.DTD	**65.43**	68.17

Table 1: WebNLG and E2ENLG Data-to-Text BLEUs

4 Conclusion

In this work, we unify encoder and decoder for both structural data and texts so as to enable a single pretraining-finetuning framework for both data and texts. Experimentally mixed data and text denoising pretraining can significantly improve data-to-text performance. Our results indicate that it is worth pursuing thorough understanding regarding how data and texts interact in denoising pretraining and finetuning.

[1] mbart50.nn is downloaded from https://github.com/pytorch/fairseq/tree/master/examples/multilingual#mbart50-models

References

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020. KGPT: knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8635–8648. Association for Computational Linguistics.

Mariana Clare, Omar Jamil, and Cyril Morcrette. 2021. A computationally efficient neural network for predicting weather forecast probabilities. *CoRR*, abs/2103.14430.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186. Association for Computational Linguistics.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Comput. Speech Lang.*, 59:123–156.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12911–12919. AAAI Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Anastasia Shimorina and Claire Gardent. 2018. Handling rare items in data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 360–370. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J. Kim. 2019. Graph transformer networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11960–11970.