2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-7281-8470-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLA51294.2020.00163

An Investigation of Fine-tuning Pre-trained Model for MR-to-Text Generation

1st Ting Hu
Internet Technologies and Systems
Hasso Plattner Institute
Potsdam, Germany
ting.hu@hpi.de

2nd Christoph Meinel
Internet Technologies and Systems
Hasso Plattner Institute
Potsdam, Germany
meinel@hpi.de

***Abstract*—Natural Language Generation (NLG) task is to generate natural language given structured data, like Meaning Representations (MRs). The generated sentences are supposed to convey all the information in MRs and be fluent and realistic as human written. The constantly emerging pre-trained models push the state-of-the-art performance of various Natural Language Processing tasks to a new level. The common practice of fine-tuning pre-trained language models for NLG is to regard it as a text-to-text generation task. That is, MRs are converted into graph structures, then the graphs are linearized as text sequences, during which lots of pre-processing and post-processing work are required. We explore different methods to organize the MRs and show that just linearizing the information in MRs achieve decent results, while complex annotation process can be omitted. Under the circumstances, further experiments demonstrate that the fine-tuned pre-trained model achieves comparable results with state-of-the-art models on the RNNLG dataset.**

***Index Terms*—Natural Language Generation, Pre-trained Model, Meaning Representation**

I. Introduction

Natural Language Generation (NLG) is the process of converting structured data, like Mean Representations (MRs), into meaningful sentences. An MR is an unordered set of attribute(slot)-value pairs, where attribute is a string and value is a sequence of words. For instance, in the restaurant NLG domain, an MR and the corresponding reference are as follow.

MR: *name[Zizzi], eatType[coffee shop], area[riverside].*

Reference: *The Zizzi is a coffee shop along the river.*

In the MR, *name[Zizzi]* is a typical slot[value] pair. NLG systems are supposed to produce utterances that exactly cover all the information provided in MRs and are influent and coherent like human-written.

Data-driven sequence-to-sequence models and their variants [1]–[3] are the mainstream on NLG task. These models are capable of generating influent and diverse utterances, while large amounts of data are required for training. In the case of insufficient data, strategies like data augmentation are employed, or else they may even be surpassed by statistical NLG systems [4].

Recently sprung up pre-trained models [5]–[8] with powerful reasoning and generation ability provide us with new possibilities. In this work, we explore fine-tuning pre-trained conditional language model BART [7] for NLG. BART is a Transformer based denoising autoencoder for pre-trained sequence-to-sequence models. In general, the structured data can be regarded as corrupted texts and used as the input of the encoder, and the decoder reconstructs original texts, that is, generating coherent sentences covering all given information in structured data.

When fine-tuning pre-trained models for NLG, the common practice is to convert the structured data into graph structures and use linearized graphs as the input of the encoder, where lots of preprocessing and postprocessing procedures are required. We investigate different methods of organizing MRs, and experiments show that just serializing the slot-value pairs in MRs achieves the best performance. In other words, the complicated annotation process is not necessary. Further experiments demonstrate that the fine-tuned BART model achieves comparable results with state-of-the-art NLG systems.

II. Related work

Natural Language Generation (NLG). NLG task varies in two formats, Knowledge Graph(KG)-to-text generation and Meaning Representation(MR)-to-text generation [9]. KG-to-text is to generate utterances given KGs, for example, WebNLG 2017 challenge [10] and DART dataset [11], where the structured data are provided in the format of triple sets. MR-to-text is to generate texts given MRs, such as E2E NLG challenge [12] and RNNLG dataset [13], which is within the scope of discussion of this paper. According to common practice, in order to convert MRs into graph structures, additional processing are essential.

Pre-trained models. Transformer based models pre-trained by massive amounts of data spring up in NLP area, such as BERT [5] and GPT-2 [6]. Among them, there are two models pre-trained in a text-to-text manner that can be employed for the NLG task. T5 [8] is pre-trained by converting different types of NLP problems into the text-to-text problem and achieves state-of-the-art performance on the CNN/Daily Mail abstractive summarization task. BART [7] is a denoising autoencoder consisting of a bidirectional encoder BERT and an auto-regressive decoder GPT-2. During pretraining, the input of the encoder is corrupted data and the decoder is supposed

978-1-7281-8470-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ICMLA51294.2020.00163

Authorized licensed use limited to: UNIVERSITY OF NEW MEXICO. Downloaded on May 15,2021 at 02:37:55 UTC from IEEE Xplore. Restrictions apply.

to reconstruct the original data. BART achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks. Considering that when obtaining comparable results on the abstractive summarization task, the size of BART model is smaller than that of T5, we employ BART for our NLG task in this work.

Fine-tuning pre-trained models. The first work to employ a pre-trained GPT-2 model for MR-to-text generation is [14]. Then [15] and [16] respectively use pre-trained T5 and GPT-2 for different structured data-to-text generation tasks. [11] propose a large open-domain structured data record-to-text generation dataset DART, and show that fine-tuning BART model on DART facilitates the out-of-domain generalization on WebNLG 2017 dataset. [9] study fine-tuning BART and T5 for graph-to-text generation in different graph domains, and manage to investigate the possible reasons for the success of pre-trained language models on graph-to-text generation tasks. The above works either make use of yet annotated Knowledge Graphs or convert MRs into graph structures for text generation, while little work is done to explore more efficient approaches of organizing MRs for fine-tuning on pre-trained models.

III. Fine-tuning BART

We explore three approaches to convert Meaning Representations (MRs) into meaningful input of the pre-trained BART model. Take one training instance in NLG dataset for one spoken dialogue system as the example below.

Dialogue Act: *inform(name=ananda fuara;pricerange=cheap ;goodformeal=lunch)*.

Reference: *ananda fuara is in the cheap price range and it is good for lunch.*

In the instance, string *inform* is a sort of Dialogue Act (DA) indicating the intention of the reference in the dialogue system, and the content in the following parentheses is the MR, where three slot-value pairs are provided.

Intuitively, the straight-forward method of aligning MRs is to linearly list the string of DA and slot-value pairs, as shown in Table I, where *source* denotes the input of encoder and *target* denotes the input of the decoder. When it comes to generating stage, only the *source* is provided and the decoder is supposed to auto-regressively generate utterances. Considering that DA strings are not required to be covered in the generation results but have an impact on the pattern of sentences, we use upper cases to distinguish the strings of DAs from slot-value pairs.

According to general seq-to-seq NLG systems, the second approach is to jointly delexicalise the MRs and references, and serialize the delexicalised slot-value pairs, displayed as method II in Table I. "delexicalise" means that specific parts of the string of MRs and references are substituted by their corresponding slots' name. Then, NLG systems are supposed to predict the pattern of the sentences rather than predict concrete content at some positions. After generation, relexicalising is employed to obtain the original natural language.

The most common method to organize structured data on NLG task is to convert them as the graph structures, and linearize them into triples [11], during which delexicalising is also indispensable. For the corresponding source format of method III in Table I, every three elements compose a triple, for example, *NAME* is the head node, *PRICERANGE* is the tail node, and *pricerange* denotes the relationship between them. Since *NAME* almost appears in all the MRs of the dataset, it is taken as the subject and regarded as the head node of each triple.

For the latter two methods, we are expecting promising results, since the task setting is to let the model learn to predict the pattern of sentences, similar to world knowledge, which better leverages the powerful reasoning capability of pre-trained BART model, while the experiment results tell us different stories.

IV. Experiment

A. Dataset

Experiments are conducted on the RNNLG dataset [13] with Dialogue Act(DA)-text pairs in Laptop, TV, Restaurant, and Hotel domains. The statistics of data in these domains are displayed in Table II. In each domain, there are different types of DAs that may influence the pattern of the generated sentences. For example, for DA *goodbye*, the corresponding reference is *thanks for visiting. goodbye for now.*, which is different from the example *inform* we show above. Hence, we also list the number of distinct DA categories and slot categories in Table II. Additionally, for DAs like *goodbye*, the number of training instances is so small that data-driven seq-to-seq NLG models may fail to show their advantages and to achieve excellent performance.

	Restaurant	Hotel	Laptop	TV
Training set	3114	3223	7994	4221
Development set	1075	547	2649	1407
Test set	1039	1075	2649	1407
DA categories	8	8	14	14
Slot categories	12	12	20	16

TABLE II
THE STATISTICS OF DATA IN FOUR DOMAINS OF RNNLG DATASET.

B. Experiment setup

We use the large version of the BART model with 12 self-attention layers in both encoder and decoder resulting in 400M parameters [17]. We use AdamW optimizer with an initial learning rate of 1×10^{-5}. The model is fine-tuned for at most 20 epochs. When decoding, beam search is employed with beam size to be 6.

C. Evaluation metrics

We use evaluations metrics BLEU and ERR[1]. ERR is computed by the number of wrongly covered slots divided by the total number of slots in input MRs. Since BLEU merely

[1]The tool is provided by https://github.com/shawnwun/RNNLG.

Authorized licensed use limited to: UNIVERSITY OF NEW MEXICO. Downloaded on May 15,2021 at 02:37:55 UTC from IEEE Xplore. Restrictions apply.

	Aligned training data
I: List slot-value pairs	source: INFORM name ananda fuara pricerange cheap goodformeal lunch target: ananda fuara is in the cheap price range and it is good for lunch.
II: Serialize delexed slot-value pairs	source: INFORM name NAME pricerange PRICERANGE goodformeal GOODFORMEAL target: NAME is in the PRICERANGE price range and it is GOODFORMEAL.
III: Linearize graph structure	source: INFORM NAME pricerange PRICERANGE NAME goodformeal GOODFORMEAL target: NAME is in the PRICERANGE price range and it is GOODFORMEAL.

TABLE I
THREE DIFFERENT METHODS OF ALIGNING MEANING REPRESENTATIONS. WHEN FINE-TUNING, SOURCE IS THE INPUT OF ENCODER AND TARGET IS THE INPUT OF THE DECODER.

Restaurant	DA: inform(name=forbes island; area=hayes valley; **kidsallowed=no**) Ref: forbes island is a nice restaurant in the hayes valley area no kid -s are allowed.
	I: forbes island is in hayes valley and does **not allow kid -s**. II: forbes island is in hayes valley and allows child -s. III: forbes island is in the hayes valley area and allows child -s.
Hotel	DA: inform_count(type=hotel; count=182; **dogsallowed=dontcare**) Ref: there are 182 hotel -s if you do not care whether dogs are allowed.
	I: there are 182 hotel -s if you **do not care whether dogs are allowed**. II: there are 182 hotel -s that allow dogs. III: there are 182 hotel -s that allow dogs.
Laptop	DA: inform_count(count=9; type=laptop; **driverange=dontcare**; pricerange=moderate) Ref: there are 9 laptop in the moderate price range. if you have no preference in the drive range.
	I: there are 9 moderate -ly priced laptop -s if you **do not care about the drive range**. II: there are 9 laptop -s in the moderate price range with dontcare drive -s. III: there are 9 laptop -s in the moderate price range with a dontcare drive.
TV	DA: recommend(name=chronos 52; type=television; audio=nicam stereo; **hasusbport=false**; resolution=1080p) Ref: the chronos 52 television has nicam stereo audio and 1080p resolution .
	I: the chronos 52 is a great 1080p television with nicam stereo audio and **no usb ports**. II: the chronos 52 is a great television with usb port -s, 1080p resolution, and nicam stereo audio. III: the chronos 52 television has usb port -s, 1080p resolution, and nicam stereo.

TABLE III
FOUR GROUPS OF GENERATED SENTENCES FROM THREE APPROACHES OF ALIGNING DATA, RESPECTIVELY.

		BLEU↑	ERR↓	MoverScore↑	BERTScore↑	BLEURT↑
Restaurant	SCLSTM	0.7525	**0.38%**	0.38	0.91	0.05
	TGen	0.7511	0.84%	0.38	0.91	0.14
	finetuned-BART	**0.7583**	0.96%	0.38	0.91	**0.21**
Hotel	SCLSTM	0.8482	3.07%	0.34	0.91	0.05
	TGen	0.8532	4.14%	0.34	0.90	0.04
	finetuned-BART	**0.8853**	**0.29%**	0.34	0.91	**0.16**
Laptop	SCLSTM	0.5116	0.79%	0.67	0.94	0.49
	TGen	**0.5150**	0.87%	0.67	0.94	0.47
	finetuned-BART	0.5133	**0.39%**	0.67	**0.95**	**0.51**
TV	SCLSTM	**0.5265**	2.31%	0.67	0.94	**0.39**
	TGen	0.5213	2.32%	0.67	0.94	0.37
	finetuned-BART	0.4213	**1.12%**	0.67	0.93	0.36

TABLE IV
EVALUATION RESULTS ON FOUR DOMAINS WHEN METHOD I IN TABLE I IS APPLIED.

evaluates word-level n-gram overlapping, recently proposed metrics BERTScore [18], MoverScore [19], and BLEURT [20] are also employed to evaluate the similarity between the references and generated candidates. BERTScore computes a similarity score for each token in the candidate with each token in the reference through contextual embeddings. MoverScore computes word mover's distance between the contextualized representations of the candidate and the reference. BLEURT measures how much information in the reference is conveyed by the candidate by applying a classification head on BERT model. These metrics demonstrate more correlation with human judgements and stronger generalization capabilities than traditional metrics.

D. Results and analysis

Experiments results are displayed in Table V. Compared with serializing delexicalised slot-values pairs and linearizing graph structure, just linearly list slot-value pairs result in the lowest ERR indicating that given structured data are best conveyed in the generated sentences. For the latter two methods, relexicalising is required after generating. Once the wrong slot names are predicted, relexicalising will fail at these postions. Such accumulated errors result in the worse performance on ERR.

Furthermore, We show four groups of generated sentences from three approaches in Restaurant, Hotel, Laptop and TV domains respectively in Table III. For the first and fourth

Authorized licensed use limited to: UNIVERSITY OF NEW MEXICO. Downloaded on May 15,2021 at 02:37:55 UTC from IEEE Xplore. Restrictions apply.

	BLEU↑	ERR↓
I: List slot-value pairs	0.5468	**0.77%**
II: Serialize delexed slot-value pairs	**0.5493**	1.02%
III: Linearize graph structure	0.5445	1.44%

TABLE V
AVERAGE EVALUATION METRICS SCORES OF ALL FOUR DOMAINS OF RNNLG DATASET WHEN THREE METHODS OF ALIGNING MRS ARE APPLIED.

group, only the generated sentences in first method correctly cover the binary slot-value pairs "kidsallowed=no" and "hasusbport=false", and the other results are just reverse. In the second and third group, only the first method of aligning data results in coheret sentences with the coverage of "dontcare" slot.

Our analysis is that for binary slots like "kidsallowed", the corresponding value with be delexicalised as "KIDSALLOWED". Then the real value of "yes" or "no" is discarded, which results in the failure of the BART model to reconstruct real data, that is, to generate sentences accurately covering the information of slot-value pairs. For neutral values like "dontcare", there are multiple methods to express the information, however, the value is delexicalised as its corresponding slot which is then regarded as an ordinary slot, thereby resulting in the lack of coherence and influence of generated sentences. Therefore, these sorts of incorrect coverage can be attributed to the inappropriate delexicalising on binary slots and neutral values.

We also compare the generation results with state-of-the-art model SCLSTM [4] and TGen [1] in Table IV. SCLSTM is a statistical language generator based on a semantically controlled LSTM structure. TGen is based on the seq2seq model with attention. According to the table, the fine-tuned BART model outperforms SCLSTM and TGen on three data domains in terms of ERR, and BLEURT, respectively. The comparable results on BERTScore and MoverScore demonstrate that the fine-tuned BART model generates candidate sentences as similar to the references as SCLSTM and TGen. Considering that our results are achieved by multi-domain training and other models are trained for each domain separately, to fine-tune BART for NLG is beneficial for practical application.

V. CONCLUSION

We have explored different methods of aligning structured data to fine-tune pre-trained denoising autoencoder BART model for NLG. Experiments on RNNLG dataset show that just linearly listing slot-value pairs in MRs as the input achieves the best performance. That is, the common practice of complicated preprocessing and postprocessing procedures, such as delexicalising and relexicalising, can be omitted under the circumstances. The further evaluation demonstrates that fine-tuned BART achieves comparable results with state-of-the-art NLG models, which are even surpassed on several data domains. For future work, we would like to investigate how more complex NLG tasks, like Knowledge Graph-to-text generation, will benefit from fine-tuning on pre-trained models.

REFERENCES

[1] O. Dušek and F. Jurčíček, "Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings," *arXiv preprint arXiv:1606.05491*, 2016.

[2] J. Juraska, P. Karagiannis, K. K. Bowden, and M. A. Walker, "A deep ensemble model with slot alignment for sequence-to-sequence natural language generation," *arXiv preprint arXiv:1805.06553*, 2018.

[3] S. Gehrmann, F. Z. Dai, H. Elder, and A. M. Rush, "End-to-end content and plan selection for data-to-text generation," *arXiv preprint arXiv:1810.04700*, 2018.

[4] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," *arXiv preprint arXiv:1508.01745*, 2015.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[8] 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," *arXiv preprint arXiv:1910.10683*, 2019.

[9] L. F. Ribeiro, M. Schmitt, H. Schütze, and I. Gurevych, "Investigating pretrained language models for graph-to-text generation," *arXiv preprint arXiv:2007.08426*, 2020.

[10] C. Gardent, A. Shimorina, S. Narayan, and L. Perez-Beltrachini, "Creating training corpora for NLG micro-planners," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 179–188. [Online]. Available: https://www.aclweb.org/anthology/P17-1017.pdf

[11] D. Radev, R. Zhang, A. Rau, A. Sivaprasad, C. Hsieh, N. F. Rajani, X. Tang, A. Vyas, N. Verma, P. Krishna *et al.*, "Dart: Open-domain structured data record to text generation," *arXiv preprint arXiv:2007.02871*, 2020.

[12] O. Dušek, J. Novikova, and V. Rieser, "Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge," *Computer Speech & Language*, vol. 59, pp. 123–156, 2020.

[13] T.-H. Wen, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, and S. Young, "Multi-domain neural network language generation for spoken dialogue systems," *arXiv preprint arXiv:1603.01232*, 2016.

[14] M. Mager, R. F. Astudillo, T. Naseem, M. A. Sultan, Y.-S. Lee, R. Florian, and S. Roukos, "Gpt-too: A language-model-first approach for amr-to-text generation," *arXiv preprint arXiv:2005.09123*, 2020.

[15] M. Kale, "Text-to-text pre-training for data-to-text tasks," *arXiv preprint arXiv:2005.10433*, 2020.

[16] H. Harkous, I. Groves, and A. Saffari, "Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity," *arXiv preprint arXiv:2004.06577*, 2020.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[18] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[19] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance," *arXiv preprint arXiv:1909.02622*, 2019.

[20] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

Authorized licensed use limited to: UNIVERSITY OF NEW MEXICO. Downloaded on May 15,2021 at 02:37:55 UTC from IEEE Xplore. Restrictions apply.