NILC at WebNLG+: Pretrained Sequence-to-Sequence Models on RDF-to-Text Generation

Marco A. Sobrevilla Cabezudo
msobrevillac@usp.br

Thiago A. S. Pardo
taspardo@icmc.usp.br

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Avenida Trabalhador São-carlense, 400. São Carlos - SP - Brazil

Abstract

This paper describes the submission by the NILC Computational Linguistics research group of the University of São Paulo/Brazil to the RDF-to-Text task for English at the WebNLG+ challenge. The success of the current pretrained models like BERT or GPT-2 in text-to-text generation tasks is well-known, however, its application/success on data-to-text generation has not been well-studied and proven. This way, we explore how good a pretrained model, in particular BART, performs on the data-to-text generation task. The results obtained were worse than the baseline and other systems in almost all automatic measures. However, the human evaluation shows better results for our system. Besides, results suggest that BART may generate paraphrases of reference texts.

1 Introduction

In recent years, the RDF-to-text generation task has gained interest from many researchers (Bouayad-Agha et al., 2014) as the RDF language, in which DBpedia is encoded, is widely used within the Linked Data framework and large scale datasets are encoded in this language.

In order to foster the use of the RDF language in this context, the first WebNLG challenge was proposed (Gardent et al., 2017). This challenge only focused on the text generation task for English and provided a test set that comprised categories included in the training set (seen categories) and categories not included in the training set (unseen categories) to evaluate the ability to generalise of the different approaches.

In general, several approaches have been explored in this task and pipeline approaches have shown a better performance than End-to-End approaches for unseen categories but not for seen ones (Castro Ferreira et al., 2019), leaving the ability to adequately deal with both categories as an open problem.

Transfer learning has gained relevance in the Natural Language Processing area and pretrained architectures like BERT (Devlin et al., 2019) or GPT (Radford et al., 2018) have outperformed prior state-of-the-art in several tasks and shown a good generalisation ability.

Even though pretrained models have been widely used in text-to-text generation (such as text simplification and automatic summarisation), this is not the case for data-to-text generation, as the input of this task is generally a graph instead of a text.

Recently, Mager et al. (2020) fine-tuned GPT-2 for a data-to-text generation task, showing improvements and that current pretrained models can deal with these representations even if the knowledge is not explicitly structured.

In this context, this paper presents the system description submitted by the NILC team to the WebNLG+ challenge 2020 (Castro-Ferreira et al., 2020). Specifically, we fine-tune BART (Lewis et al., 2020), a denoising autoencoder for pretraining sequence-to-sequence models, on the RDF-to-text generation dataset provided by this task.[1]

2 WebNLG+ Challenge

The first WebNLG challenge (Gardent et al., 2017) consisted in generating English text from a set of RDF triples extracted from DBpedia. Differently from the previous edition, this edition comprises two tasks:

- RDF-to-text generation, similarly to WebNLG 2017 but with new data and for English and Russian;
- Text-to-RDF semantic parsing: converting a text into the corresponding set of RDF triples.

[1] The corresponding source code is available at https://github.com/msobrevillac/webnlg-2020-bart

3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), Dublin, Ireland (Virtual), 18 December 2020, pages 131–136, *©2020 Association for Computational Linguistics*
Attribution 4.0 International.

Figure 1 shows an example of the triples-text pair. In particular, the RDF-to-text generation involves NLG subtasks such as Discourse ordering (how to order the RDF triples), Text Structuring (how to cluster triples in sentences), lexicalisation (how to find the proper phrases and words to express the content to be included in each sentence), Referring Expression Generation (how to generate the references to the entities of the discourse), and surface realisation (how to convert non-linguistic data into text).

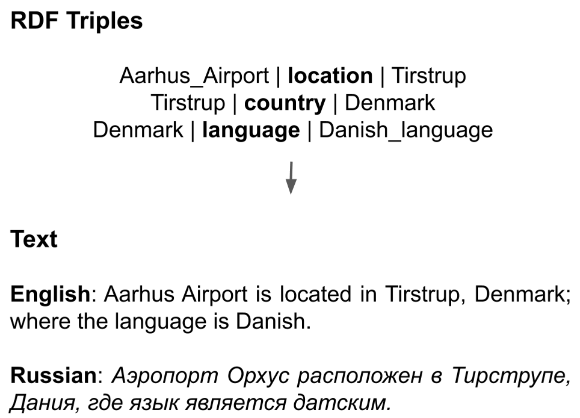

Figure 1: Example of a set of triples (top) and the corresponding text in English and Russian (bottom).

It is worth noting that the WebNLG dataset comprises different categories (domains) and the test set comprises instances belonging to the categories included in the training set and instances belonging to new unseen categories. Some instances also contain entities not seen in the training set.

3 BART

BART (Lewis et al., 2020) is a denoising autoencoder for pretraining sequence-to-sequence models. It is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Transformer-based neural machine translation architecture (Vaswani et al., 2017) with a bidirectional encoder similar to BERT (Devlin et al., 2019) and a left-to-right decoder similar to GPT (Radford et al., 2018) (Figure 2).

4 System Description

As mentioned, our approach is based on BART. Thus, in order to preprocess the input for BART, we linearise the triples by putting all triples (in

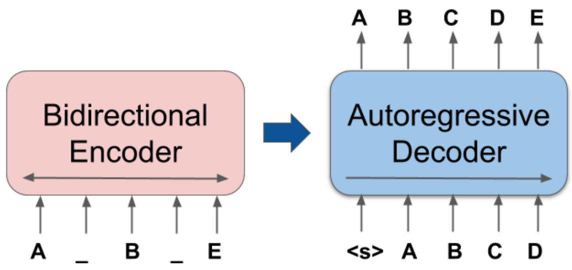

Figure 2: BART architecture. Extracted from (Lewis et al., 2020).

the form entity, relation and entity/expression) sequentially. We also remove underscores from the entities and expressions, split relations according to uppercase tokens, remove quotes from the expressions and put all in lowercase. For example, the triple "*Aarhus_Airport* ***operatingOrganisation*** "*Aarhus Lufthavn AS*"" is converted into "*aarhus airport operating organisation aarhus lufthavn a/s*". It is worth noting that we tried other linearisation strategy by putting a dot mark between each preprocessed triple, but this alternative did not lead to improvements.

Additionally, we train a model to recase the sentences by using the tool provided by Moses[2] and, finally, we tokenise and convert the sentences to lowercase.

We use the large BART model provided by HuggingFace (Wolf et al., 2019). We finetune the model for 5 epochs, using a batch size of 16, the Adam optimiser with a learning rate of 0.0001, a max length of 100 in the source and target. For the decoding, we use a beam size of 5.

For the post-processing, we use the recaser previously trained, normalise the punctuation and detokenise the outputs[3].

5 Results and Discussion

The performance of the several proposals at the challenge is computed by using BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), TER (Snover et al., 2006), chrF++ (Popović, 2017), Bertscore (Zhang* et al., 2020), and BLEURT (Sellam et al., 2020).[4]

[2] Available at https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/train-recaser.perl

[3] We use the perl code available at https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer for the punctuation normaliser and the detokeniser.

[4] The platform used to compute the measures is the one proposed by Moussalem et al. (2020).

Table 1 shows the results of our approach and the baselines.[5] In general, our approach obtained worse performances than the baselines, only outperforming both baselines for seen domains. A possible explanation to these results is related to the embeddings, as we did not freeze any layer during training and it could affect the performance on unseen categories.

Other explanation is the way as we linearise the RDF triples, as BART could not distinguish the different triples in the input. In recent work, Ribeiro et al. (2020) show similar results to ours on seen categories. Besides, they show that performance on RDF-to-text generation is less influenced by the order of the input than other representations. However, a deeper study is necessary to explore the performance on unseen domains.

A point to highlight is that even though our results were not good enough, when comparing with approaches that got similar results, we may see that METEOR and BLEURT produce better results (Table 2).[6] We hypothesise that, in some cases, BART generates paraphrases of the correct sentences. This way, we got better results for METEOR and BLEURT (and BERTScore) because these metrics are more related to semantics instead of n-gram overlapping (like BLEU, TER, or other measures).

Table 4 shows the overall results of the human evaluation for our system and the three systems used in Table 2.[7] The organisers aimed to measure the following criteria:

- Data coverage: this criterion assesses how much information from the data has been covered in the text;
- Relevance: this criterion evaluates if the text contains any non-presented predicates;
- Correctness: annotators were asked to evaluate if the text describes predicates (which are both in data and text) with correct objects. The subject must also be correctly described;
- Text structure: this criterion evaluates if the text is grammatical and well-structured, written in good English; and
- Fluency: the annotators were asked to evaluate if the text progresses naturally and sounds like a coherent unit.

Each criterion is rated with a single number in the range from ”0” (completely disagree) to ”100” (completely agree). The scores as they appear for each criterion have been normalised (z-scores) and clustered into groups among which there are no statistically significant differences according to the Wilcoxon rank-sum significant test.

In general, our system was ranked at 3rd and 4th cluster, except for fluency in which it was ranked at the last cluster. It is worth noting that even though some metrics like BLEU or chrF++ showed that our system performed worse than the other ones (UPC-POE and ORANGE-NLG teams), results in the human evaluation were opposite and seemed to be more correlated with metrics like METEOR or BLEURT. Furthermore, our system got similar results (without statistical significance differences) to the one proposed by Huawei even when the results in automatic evaluation showed a lower performance for our system. All these results reinforce the idea that our proposal could be generating paraphrases in the output.

Other point to highlight is the result obtained in fluency. We expected that our approach would get better results as it was trained on large corpora and this kind of pretrained models tend to get good performance in terms of fluency.

Finally, Table 3 shows the results of the human evaluation for our system on seen domains, unseen domains, and unseen entities. As it can be seen, our system performs well on seen domains but not on unseen domains and entities (ranked at the last cluster). This result could have been produced by problems in the embeddings as we did not freeze these at training time.

6 Conclusion and Future Work

This paper described the application of a pretrained sequence-to-sequence model, called BART, to the RDF-to-text generation task in the context of the WebNLG+ challenge. Results suggest that BART generates paraphrases of the reference text, as evaluation metrics more related to semantics got better

[5] The approach of the baseline in Table 1 has not been revealed at the time of this paper version. The approach of Baseline 2 is one proposed by Gardent et al. (2017), which is based on Neural Machine Translation and delexicalisation.

[6] All results are available at https://beng.dice-research.org/gerbil/webnlg2020results.

[7] All results are available at https://beng.dice-research.org/gerbil/webnlg2020resultshumaneval.

	BLEU	METEOR	chrF++	TER	BERT-F1	BLEURT
Test - All						
Baseline	40.57	0.373	0.621	0.517	0.943	0.47
Baseline 2	37.89	0.364	0.606	0.553	0.930	0.42
Our approach	31.98	0.350	0.545	0.629	0.920	0.40
Test - Seen Domains						
Baseline	42.95	0.387	0.650	0.563	0.943	0.41
Baseline 2	41.15	0.384	0.642	0.599	0.936	0.33
Our approach	56.18	0.409	0.700	0.430	0.958	0.58
Test - Unseen Domains						
Baseline	37.56	0.357	0.584	0.510	0.940	0.44
Baseline 2	34.63	0.347	0.565	0.544	0.925	0.39
Our approach	16.2	0.311	0.435	0.719	0.902	0.19
Test - Unseen Entities						
Baseline	40.22	0.384	0.648	0.476	0.949	0.55
Baseline 2	38.07	0.367	0.626	0.515	0.932	0.50
Our approach	21.93	0.340	0.509	0.671	0.916	0.42

Table 1: Results of our system and the baselines on test set.

	BLEU	METEOR	chrF++	BERT-F1	BLEURT
Ours	31.98	0.350	0.545	0.920	0.40
UPC-POE / id14	39.12	0.337	0.579	0.929	0.37
ORANGE-NLG / id13	38.20	0.335	0.571	0.920	0.29
Huawei / id17	39.55	0.372	0.613	0.935	0.37

Table 2: Results of our system and some approaches with similar results.

System	Rank	Avg. Z	Avg. Raw
Data coverage			
Ours	4/6	-0.477	81.605
UPC-POE / id14	6/6	-0.782	75.845
ORANGE-NLG / id13	5/6	-0.554	79.959
Huawei / id17	4/6	-0.31	84.743
Relevance			
Ours	3/4	-0.499	83.522
UPC-POE / id14	4/4	-0.531	82.051
ORANGE-NLG / id13	4/4	-0.71	79.887
Huawei / id17	3/4	-0.425	85.265
Correctness			
Ours	3/4	-0.589	76.702
UPC-POE / id14	4/4	-0.701	74.374
ORANGE-NLG // id13	4/4	-0.668	74.977
Huawei / id17	3/4	-0.389	80.76
Text structure			
Ours	3/4	-0.402	80.463
UPC-POE / id14	4/4	-0.456	78.503
ORANGE-NLG / id13	3/4	-0.338	80.462
Huawei / id17	3/4	-0.373	80.219
Fluency			
Ours	5/5	-0.408	74.851
UPC-POE / id14	5/5	-0.508	72.28
ORANGE-NLG / id13	5/5	-0.332	75.675
Huawei / id17	5/5	-0.369	75.205

Table 3: Results of the human evaluation for our system and some approaches with similar results.

results than the ones that are more related to n-gram overlapping.

As future work, we plan to evaluate other alternatives for the linearisation process and use multilingual BART (Liu et al., 2020) in order to deal with the same task for Russian.

Acknowledgements

This work was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)* – Finance Code 88882.328822/2019-01. The authors are also grateful to USP Research Office (PRP 668) for supporting this work, and would like to thank NVIDIA for donating the GPU. This work is part of the OPINANDO project (`https://sites.google.com/icmc.usp.br/opinando/`) and the USP/FAPESP/IBM Center for Artificial Intelligence (C4AI - `http://c4ai.inova.usp.br/`). Finally, this research is carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0).

References

Nadjet Bouayad-Agha, Gerard Casamayor, and Leo Wanner. 2014. Natural language generation in the context of the semantic web. *Semantic Web*, 5(6):493–513.

Thiago Castro-Ferreira, Claire Gardent, Nikolai

	Data Coverage		Relevance		Correctness		Text Structure		Fluency	
	Rank	Avg. Z	Rank	Avg. Z	Rank	Avg. Z	Rank	Avg. Z	Rank	Avg. Z
Seen Domains	1/3	0.225	1/2	0.266	1/2	0.212	1/3	0.212	2/3	0.155
Unseen Domains	5/5	-0.97	4/4	-1.06	4/4	-1.098	4/4	-0.685	4/4	-0.721
Unseen Entities	3/4	-0.343	3/3	-0.299	2/3	-0.563	3/3	-0.629	3/3	-0.492

Table 4: Results of the human evaluation for our system and some approaches with similar results.

Ilinykh, Chris van der Lee, Simon Mille, Diego Moussalem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional webnlg+ shared task: Overview and evaluation results (webnlg+ 2020). In *Proceedings of the 3rd WebNLG Workshop on Natural Language Generation from the Semantic Web (WebNLG+ 2020)*, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

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*, pages 7871–7880, Online. 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.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1846–1852, Online. Association for Computational Linguistics.

Diego Moussalem, Paramjot Kaur, Thiago Castro-Ferreira, Chris van der Lee, Conrads Felix Shimorina, Anastasia, Michael Röder, René Speck, Claire Gardent, Simon Mille, Nikolai Ilinykh, and Axel-Cyrille Ngonga Ngomo. 2020. A general benchmarking framework for text generation. In *Proceedings of the 3rd WebNLG Workshop on Natural Language Generation from the Semantic Web (WebNLG+ 2020)*, Dublin, Ireland (Virtual). Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, A. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.