WebNLG-Interno: Utilizing FRED-T5 to address the RDF-to-text problem

Maxim Kazakov
Petal Cloud Technology Co.,Ltd
HSE University
kazakov.maxim@huawei.com

Julia Preobrazhenskaya
Lobachevsky State University
enjulia17@gmail.com

Ivan Bulychev
Lobachevsky State University
halkimic@gmail.com

Aleksandr Shain
Petal Cloud Technology Co.,Ltd
Pulkovo Observatory
alexander.shain@huawei.com

Abstract

We present our solution for the Russian RDF-to-text generation task of the WebNLG Challenge 2023 [1]. We use the pretrained large language model named FRED-T5 (Zmitrovich et al., 2023) to finetune on the train dataset. Also, we propose several types of prompt and run experiments to analyze their effectiveness. Our submission achieves 0.373 TER on the test dataset, taking the first place according to the results of the automatic evaluation and outperforming the best result of the previous challenge by 0.025. The code of our solution is available at the following link: https://github.com/Ivan30003/webnlg_interno

1 Introduction

Recently released large language models (a.k.a. LLMs) like GPT-4 (OpenAI, 2023), BLOOM (Scao et al., 2022), LLaMA (Touvron et al., 2023) and PaLM (Chowdhery et al., 2022) proved the ability of deep neural networks to generate realistic texts, maintain a human-like conversation and answer factual questions based on the information contained in the training data. However, real-life applications of LLMs could benefit from extracting relevant information from external databases to provide a user with a proper answer. Data in databases typically have a structured form, and one of the main challenges is to present extracted information in the natural sentences.

The RDF-to-text track of the WebNLG challenge aims to address that particular problem. Given the data presented in a common RDF format, the task is to generate a natural utterance that conveys information from the structured data. The RDF data format operates with three entities - subject, object and predicate, where the latter represents a type of relations between subject and object. Such entities form a triple, and each data sample is represented as a tripleset consisting of one or several triples (Table 1).

It was mentioned by Kasner and Dušek (2020), such tripleset representation in English can be seen as a noisy version of a target utterance, and denoising autoencoders like T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) may provide a stable solution to the problem. In Russian track of the challenge we additionally expect generated utterances to be in Russian while the triples remain in English. That can be handled either by a two stage solution where translation (of input triples or generated utterances) performed as a separate step, or by an end-to-end solution - for the such case, encoder-decoder architectures are suitable as well (Liu et al., 2020).

The above assumptions are supported by the results of the previous challenge (Castro Ferreira et al., 2020), where 3 out of 6 solutions for the RDF-to-text Russian track were based on a pretrained multilingual BART model (Kasner and Dušek, 2020; Yang et al., 2020; Li et al., 2020), and the winning solution was based on a T5 model pretrained on a large bilingual (*en*, *ru*) corpus of structured data (Agarwal et al., 2020).

In contrast to the previous solutions, we focus solely on the Russian track and use the state-of-the-art LLM for russian language named FRED-T5 (Zmitrovich et al., 2023), assuming that pretraning on a large corpus of Russian texts can benefit the model's ability to generate more realistic utterances. We also experiment with prompt con-

[1] https://synalp.gitlabpages.inria.fr/webnlg-challenge/challenge_2023

Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023), pages 67–72
September 12, 2023

XML	<entry category="Airport" eid="1309" shape="(X (X (X)))" shape_type="chain" size="2"> < originaltripleset > <otriple>Aarhus_Airport | location | Tirstrup</otriple> <otriple>Tirstrup | country | Denmark</otriple> </ originaltripleset > <modifiedtripleset> <mtriple>Aarhus_Airport | location | Tirstrup</mtriple> <mtriple>Tirstrup | country | Denmark</mtriple> </modifiedtripleset> <lex comment="" lang="en" lid="Id1">Aarhus Airport is located in Tirstrup, Denmark.</lex> <lex comment="" lang="en" lid="Id2">The location of Aarhus Airport is Tirstrup, in Denmark.</lex> <dbpedialinks> <dbpedialink direction="en2ru">Denmark | sameAs | Дания</dbpedialink> </dbpedialinks> <links> <link direction="en2ru">Tirstrup | sameAs | Тирструп</link> <link direction="en2ru">Aarhus Airport | includes | Орхус</link> </links> </entry>
Labels	Аэропорт Орхус расположен в Тирструпе, Дания. Место расположения орхусского аэропорта - Тирструп, в Дании.

Table 1: Example of XML data and labels for Russian track of WebNLG Challenge. *<modifiedtripleset>* is used as linearized triples.

struction enriching tripleset with additional data presented in raw XML. Our final submission is based on a FRED-T5 model finetuned on the train dataset with automatically translated predicates and the properties "links", "dbpedialinks", "category" and "size" included into the prompt. Our submission achieves 0.373 TER on the test dataset, taking the first place according to the results of the automatic evaluation and outperforming the best result of the previous challenge by 0.025.

2 Data

The original dataset for RDF-to-text task is presented in the XML-file format. It is split into train, dev and test sets. RDF-triples were extracted from DBpedia and have the subject-predicate-object structure linearized with vertical bars as separators (Table 1). The number of triples in data samples varies from 1 to 7, and each sample belongs to one of 9 categories: Building, Astronaut, Airport, SportsTeam, ComicsCharacter, CelestialBody, Monument, University, Food.

Category names and triples are initially presented in English. However, links between English and Russian entities from subjects and objects of RDF-triples are provided by *<links>* and *<dbpedialinks>* properties. These links are encoded by two kinds of relations: "includes" and "sameAs", the latter of which can be considered as an explicit translation.

The model's input is the prompt, and we discuss the prompt construction in the next section. As an output, we expect to receive a generated text similar to the text from the label. However, each sample may contain multiple labels that represent a natural language text corresponding to triples (Table 1). Hence we transform each label with the corresponding input into a separate sample for training. In total, the training dataset consists of 14630 samples (from 5573 original samples). Validation and test datasets have 790 and 1102 samples, respectively, on which we performed multi-reference model evaluation.

3 Experimental setup

3.1 Triple Processing

In our approach, data preparation process included two stages - data pre-processing and prompt construction. At data pre-pocessing stage, we converted predicate names and categories from so called 'camel' case to multi-word expressions (e.g. "isPartOf" becomes "is part of"), analogous to (Agarwal et al., 2020). Also, we removed underscore from all occurrences of Object and Subject names in linearized triples, *<links>* and *<dbpedialinks>* (e.g. "Aarhus_Airport" becomes "Aarhus Airport").

At prompt construction stage, we considered three ways of constructing prompt (Table 2). In a "Simple" case, the prompt is presented as a set of pre-processed triples, extracted from the original dataset and separated by semicolon. Unlike the works presented at the challenge in previous years, we did not use any special tokens to indicate the subject, predicate and object.

However, it was noticed that the model faces the problem of translating proper nouns from English to Russian: in some cases the translation was either incorrect or sounded unnatural. To solve this issue, pre-processed links with subjects and

Name	Prompt
Simple	{ Aarhus Airport | location | Tirstrup }; { Tirstrup | country | Denmark }
With links	Соотношения: { Aarhus Airport | location | Tirstrup }; { Tirstrup | country | Denmark }.\nДополнительные соотношения: { Tirstrup = Тирструп }; { Aarhus Airport >= Орхус }.\nСсылки: { Denmark = Дания }.
Full	Категория: Airport.\nЧисло соотношений: 2. Соотношения: { Aarhus Airport | location | Tirstrup }; { Tirstrup | country | Denmark }.\nДополнительные соотношения: { Tirstrup = Тирструп }; { Aarhus Airport >= Орхус }.\nСсылки: { Denmark = Дания }.\nКороткое высказывание:
Translated	Категория: Аэропорт.\nЧисло соотношений: 2. Соотношения: { Aarhus Airport | месторасположение | Tirstrup }; { Tirstrup | страна | Denmark }.\nДополнительные соотношения: { Tirstrup = Тирструп }; { Aarhus Airport >= Орхус }.\nСсылки: { Denmark = Дания }.\nКороткое высказывание:

Table 2: Examples of the proposed prompts. Translation of keywords: Соотношения - Relations; Дополнительные соотношения - Additional relations; Ссылки - Links; Категория - Category; Число соотношений - Number of relations; Короткое высказывание - Short statement.

objects translation from the original dataset were included in the basic version. We preserved original <*links*> and <*dbpedialinks*> division by using keywords "Ссылки" and "Дополнительные соотношения", respectively, and encoded "sameAs" and "includes" relations by "=" and ">=". Moreover, we added a keyword "Соотношения" for the triples to orient a model in the given data. We denote such prompt as "With links".

In order to provide a model with the context and simplify its "perception" process of the received data, it was suggested to add some metadata, such as pre-processed category names and tripleset size, to the prompt. We denote that prompt as "Full".

To prepare our final dataset, we automatically translated predicates and category names (without taking into account the context), expecting that to reduce the number of grammar and morphological mistakes while generating sentences in Russian. The minimum number of characters in our prompt is 171, the maximum is 2179, the average is 450. Examples of each prompt construction is presented in Table 2.

3.2 Training setup

For our experiments we use a large language model named FRED-T5 with 1.7 billion parameters (Zmitrovich et al., 2023). The model is based on T5 architecture (Raffel et al., 2020) and trained on a large corpus of Russian texts using a mixture of denoising objectives analogous to UL2 (Tay et al., 2023). We also experiment with a multilingual mT5 model (Xue et al., 2021) in *Large* (1.2B) and *XL* (3.7B) configurations. Comparing the results of FRED-T5 and mT5, we aim to examine how critical it is for the task to use a model trained specifically for Russian language.

	LoRA trainable # params	Total # params
FRED-T5	7 077 888	1 747 435 008
mT5-Large	4 718 592	1 234 299 904
mT5-XL	9 437 184	3 752 056 832

Table 3: Model size

To start with, we used the pretrained checkpoints [2] [3] [4] and finetuned the models solely on the train dataset for 20 epochs with total $batch\_size = 16$ on $4 \times V100$ GPUs. To make training process more efficient, we used LoRA method (Hu et al., 2022) with $rank = 16, \alpha = 32, dropout = 0.05$ as a commonly used configuration. We used standard

[2]FRED-T5: https://huggingface.co/ai-forever/FRED-T5-1.7B

[3]mT5-Large: https://huggingface.co/google/mt5-large

[4]mT5-XL: https://huggingface.co/google/mt5-xl

Team Name	BLEU	METEOR	chrF++	TER ↓	BERT F1
WebNLG-Interno (ours)	**54.711**	**0.700**	**0.690**	**0.373**	**0.920**
cuni-ufal (Kasner and Dušek, 2020)	52.930	0.672	0.677	0.398	0.909
bt5 (Agarwal et al., 2020)	51.630	0.676	0.683	0.420	0.907
Baseline (Castro Ferreira et al., 2020)	25.500	0.467	0.514	0.665	0.837

Table 4: Automatic evaluation results of our submission and the leaders of the previous challenge.

cross-entropy loss objective with label smoothing factor $alpha = 0.1$ for model finetuning. It was decided to choose the best checkpoint by the METEOR value on the dev dataset since this metric has been used as an objective in the previous challenge (Castro Ferreira et al., 2020). To obtain the final results, we used beam search with $width = 5$.

4 Results

In this section we provide the results of our experiments. We finetuned FRED-T5 model on the training dataset with different prompts. Table 5 shows the results of automatic evaluation on dev and test splits. Comparing the results of finetuning the model on "Simple" and "With links" prompts, we observe that enriching prompt with *<links>* and *<dbpedialinks>* data leads to a significant gain in generation quality on average. However, this gain is not consistent as we noticed drop in performance for a big portion of samples. Also, we cannot confirm that using *<links>* and *<dbpedialinks>* data contributes to a better translation of named entities, although it does help for some samples. It seems that the additional data simply enables a better convergence. The same conclusion was made for the "Full" prompt, although the gain is not so noticeable.

Also, we conducted the same experiments with translated predicates and categories. To our surprise, this did not lead to a significant improvement, and for the "Simple" prompt it even worsened the model's performance. The possible explanation for this is that the target labels are quite different from the translated predicates, and FRED-T5 already has a decent translation ability.

In order to understand what advantages FRED-T5 has over multilingual LLMs, we finetuned mT5-Large and mT5-XL models on the train dataset with "Full Translated" prompt. From Table 6 we can conclude that mT5-XL and FRED-T5 demonstrate comparable performance, while FRED-T5 is more than twice smaller than mT5-XL (Table 3). At the same time, mT5-Large is closer to FRED-T5 in terms of model size, but converges with lower accuracy, especially on the dev dataset. Considering the fact that from architecture perspective all models are based on T5, we may assume that FRED-T5 benefits from its pretraining scheme and the target language corpora.

In our submission we used FRED-T5 model finetuned on the train dataset with "Full Translated" prompt. Table 4 shows automatic evaluation results on the test dataset in comparison to the leaders of the previous challenge (Castro Ferreira et al., 2020). Our model outperforms existing solutions by a significant margin, although does not yield a drastic improvement.

5 Conclusion

We presented a solution for the RDF-to-text problem in Russian. Our solution is based on FRED-T5 large language model, utilizes additional information from raw XML data and is aided by machine translation. The developed solution achieves 0.373 TER on the WebNLG-2023 test dataset and outperforms existing solutions by a large margin. Furthermore, conducted experiments demonstrated that translated data is not crucial for the solution and provides only a small gain, while a proper pretraining plays a major role.

Acknowledgements

We would like to express our gratitude to Wenshuai Yin and Alexey Trushkov for their support in our participation in the challenge, and to Leonid Beynenson for his valuable technical advices.

References

Oshin Agarwal, Mihir Kale, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Machine translation aided bilingual data-to-text generation and semantic parsing. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 125–130, Dublin,

Dev						
	Simple		With links		Full	
	Original	Translated	Original	Translated	Original	Translated
BLEU	49.338	48.286	50.356	**51.020**	50.630	50.708
METEOR	0.65	0.64	**0.66**	**0.66**	**0.66**	**0.66**
chrF++	0.66	0.65	**0.67**	**0.67**	**0.67**	**0.67**
TER ↓	0.442	0.443	0.423	0.419	0.421	**0.416**
BERT-SCORE P	0.90	0.90	**0.91**	**0.91**	**0.91**	**0.91**
BERT-SCORE R	**0.90**	0.89	**0.90**	**0.90**	**0.90**	**0.90**
BERT-SCORE F1	0.90	0.90	0.90	0.90	0.90	**0.91**

Test						
	Simple		With links		Full	
	Original	Translated	Original	Translated	Original	Translated
BLEU	54.173	54.028	54.449	**54.960**	54.956	54.711
METEOR	0.69	0.69	0.69	0.69	**0.70**	**0.70**
chrF++	**0.69**	0.68	**0.69**	**0.69**	**0.69**	**0.69**
TER ↓	0.386	0.389	0.381	0.381	0.379	**0.373**
BERT-SCORE P	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**
BERT-SCORE R	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**
BERT-SCORE F1	0.91	0.91	0.91	**0.92**	**0.92**	**0.92**

Table 5: FRED-T5 results. Best checkpoint is chosen by METEOR value on dev split. The completions obtained using beam search with $width = 5$.

Dev			
	FRED-T5	mT5 XL	mT5 Large
BLEU	**50.708**	50.362	48.588
METEOR	**0.66**	0.65	0.64
chrF++	**0.67**	0.66	0.66
TER ↓	0.416	**0.411**	0.434
BERT P	**0.91**	**0.91**	**0.91**
BERT R	**0.90**	**0.90**	**0.90**
BERT F1	**0.91**	0.90	0.90

Test			
	FRED-T5	mT5 XL	mT5 Large
BLEU	**54.711**	54.688	54.121
METEOR	**0.70**	0.69	0.69
chrF++	**0.69**	**0.69**	**0.69**
TER ↓	**0.373**	0.380	0.377
BERT P	**0.92**	**0.92**	**0.92**
BERT R	**0.91**	**0.91**	**0.91**
BERT F1	**0.92**	0.91	**0.92**

Table 6: Results of the models finetuned using the "Full Translated" prompt. Best checkpoint is chosen by METEOR value on dev split. The completions obtained using beam search with $width = 5$.

Ireland (Virtual). Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zdeněk Kasner and Ondřej Dušek. 2020. Train hard, finetune easy: Multilingual denoising for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 171–176, Dublin, Ireland (Virtual). 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.

Xintong Li, Aleksandre Maskharashvili, Symon Jory Stevens-Guille, and Michael White. 2020. Leveraging large pretrained models for WebNLG 2020. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 117–124, Dublin, Ireland (Virtual). 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.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zixiaofan Yang, Arash Einolghozati, Hakan Inan, Keith Diedrick, Angela Fan, Pinar Donmez, and Sonal Gupta. 2020. Improving text-to-text pre-trained models for the graph-to-text task. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 107–116, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Dmitry Zmitrovich, Andrei Kalmykov, Vitaly Kadulin, Mikhail Novikov, and Alexey Khoroshilov. 2023. FRED: Full-scale russian enhanced denoisers T5.