Train Hard, Finetune Easy: Multilingual Denoising for RDF-to-Text Generation

Zdeněk Kasner and Ondřej Dušek
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics,
Prague, Czech Republic
{kasner,odusek}@ufal.mff.cuni.cz

Abstract

We describe our system for the RDF-to-text generation task of the WebNLG Challenge 2020. We base our approach on the mBART model, which is pre-trained for multilingual denoising. This allows us to use a simple, identical, end-to-end setup for both English and Russian. Requiring minimal task- or language-specific effort, our model placed in the first third of the leaderboard for English and first or second for Russian on automatic metrics, and it made it into the best or second-best system cluster on human evaluation.

1 Introduction

The landscape of approaches for text generation has evolved since the first edition of the WebNLG challenge. Self-supervised *pre-training* objectives—such as language modelling and text denoising—have proven efficient for training neural models with excellent surface realization capabilities (Devlin et al., 2019; Lewis et al., 2020). Pre-training is used to improve the performance of models on downstream tasks, requiring only a small amount of task-specific data (Chen et al., 2020).

Pre-trained models can exploit shared representations across languages, following the success of multilingual word embeddings (Chen and Cardie, 2018; Lample and Conneau, 2019). Although multilingual pre-training (i.e., pre-training on a collection of corpora from multiple languages) may slightly hurt performance for high-resource languages, it allows using the models for crosslingual tasks (Liu et al., 2020; Conneau et al., 2020).

Neural architectures for text generation also gave rise to end-to-end approaches, where inputs and outputs are linearized and the task is solved by a single neural sequence-to-sequence model. Despite its disproportionate simplicity, this approach can be hard to beat using task-specific, modular approaches (Dušek et al., 2020).

In our submission, we took advantage of recent advances in pre-trained denoising autoencoders, multilingual representations, and sequence-to-sequence approaches. They enabled us to approach RDF-to-text generation both in English and Russian with a simple, identical, end-to-end setup. We finetune the pre-trained mBART model (Liu et al., 2020) on the provided training data individually for each language. We feed tokenized and trivially linearized input RDF triples into the model and train it to output ground-truth references. We do not use any additional preprocessing, postprocessing, or other intermediate steps.

Originally, this approach was just a baseline that we planned to improve. However, the baseline approach yielded results of such quality that we decided to use it for our official WebNLG submission. The results of automatic metrics (Moussalem et al., 2020)[1] and human evaluation as well as our manual inspections confirmed our expectations. In automatic metrics, our solution placed in the top third of the field (out of 35 submissions) for English and first or second (out of 12 submissions) for Russian. In human evaluation, it scored in the best or second-best system cluster. We believe that our approach—with its excessive simplicity—can serve as a benchmark for a trade-off between the output quality and the setup complexity.

2 Task Description

The WebNLG Challenge 2020 (Castro-Ferreira et al., 2020)[2] is the second edition of the shared task in mapping structured data to text. The data contains sets of RDF triples extracted from DBpedia accompanied with verbalizations which were crowdsourced from human annotators.

[1] https://gerbil-nlg.dice-research.org/gerbil/webnlg2020results

[2] https://webnlg-challenge.loria.fr/challenge_2020/

3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+), Dublin, Ireland (Virtual), 18 December 2020, pages 171–176, ©2020 Association for Computational Linguistics

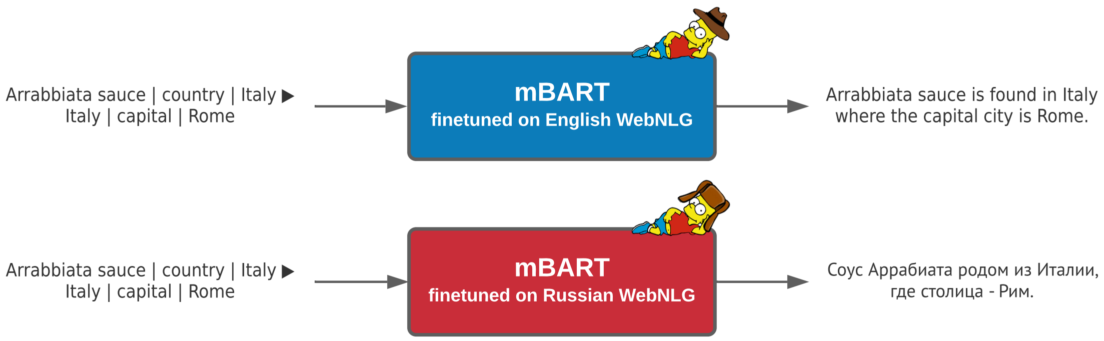
Arrabbiata sauce | country | Italy ▶
Italy | capital | Rome
mBART
finetuned on English WebNLG
Arrabbiata sauce is found in Italy
where the capital city is Rome.
Arrabbiata sauce | country | Italy ▶
Italy | capital | Rome
mBART
finetuned on Russian WebNLG
Соус Аррабиата родом из Италии,
где столица - Рим.

Figure 1: Our setup is simple: after tokenizing and linearizing the RDF triples, we finetune two separate mBART models for English and Russian using provided training data. We submit the unprocessed output from each model.

The original challenge (Gardent et al., 2017a,b) included 10 categories in the training data: *Airport, Astronaut, Building, City, ComicsCharacter, Food, Monument, SportsTeam, University*, and *WrittenWork*. Each set of triples included several verbalizations to promote lexical variability. WebNLG 2020 includes several extensions:

(1) It is *bilingual*: in addition to original English data, a new portion of the dataset with Russian lexicalizations is provided, giving rise to a new task of generating text in Russian.

(2) It is *bidirectional*: in addition to RDF-to-text generation, the challenge also includes a task on text-to-RDF semantic parsing. (We did not participate in this task.)

(3) It includes 6 *new categories*: 5 unseen categories from WebNLG Challenge 2017 (*Athlete, Artist, CelestialBody, MeanOfTransportation, Politician*) and 1 new category (*Company*).

3 Multilingual Denoising

Denoising autoencoders are trained to take a partially corrupted input and restore the original undistorted input by minimizing the reconstruction error (Vincent et al., 2010). On top of regular autoencoders, the model is forced to extract high-level features from the input distribution to filter out the noise. With a suitable noise function, denoising autoencoders can be trained in a self-supervised way on large datasets.

BART (Lewis et al., 2020) is a denoising autoencoder with an objective of restoring a corrupted document. The model uses an encoder-decoder architecture: the bi-directional encoder encodes the corrupted input; the left-to-right decoder aims to restore the original, undistorted input. The model can be seen as a generalization of both BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019).

Adopting BART's objective and architecture, mBART (Liu et al., 2020) is pre-trained on the large-scale CC25 corpus extracted from Common Crawl, which contains data in 25 languages (Wenzek et al., 2020). The data is tokenized using a SentencePiece model (Kudo and Richardson, 2018) trained on the training corpus with a vocabulary of 250,000 subword tokens. The noise function of mBART replaces text spans of arbitrary length with a mask token (35% of the words in each instance) and permutes the order of sentences. The model uses the Transformer architecture (Vaswani et al., 2017) with 12 layers for the encoder and 12 layers for the decoder (∼680M parameters).

4 Our Submission

We formulate the RDF-to-text task as *text denoising* and train mBART to solve the task individually for each language (see Figure 1). We use the provided XML WebNLG data reader[3] to load and linearize the triples. For each triple, we use the `flat_triple()` method which converts each triple into the following format:

```
subject | property | object
```

Note that the constituents of the triple (subject, predicate, object) are only marked positionally, without any extra tags. We use a token not present in the training data ("▶") for delimiting individual triples to avoid extending the model vocabulary. We linearize the triples in their default order.

[3] https://gitlab.com/webnlg/corpus-reader

input	`Piotr_Hallmann | weight | 70.308 ▶ Piotr_Hallmann | birthDate | 1987-08-25`
output [en]	Born on August 25th 1987, Piotr Hallmann has a weight of 70.308.
input	`Ciudad_Ayala | populationMetro | 1777539`
output [en]	The population metro of Ciudad Ayala is 1777539.
input	`Bakewell_tart | ingredient | Frangipane`
output [ru]	Франжипан - один из ингредиентов тарта Бейквелл.
transcription	Franzhipan - odin iz ingredientov tarta Bejkvell.
translation	Frangipane is one of the ingredients of the Bakewell tart.

Table 1: Example outputs from the mBART model(s) finetuned for RDF-to-text generation. (1) The model can work with unseen entities, dates and numbers. (2) The model is quite robust to unseen properties, such as `populationMetro`. However, the surface form of the property deviates too much from its meaning and the sentence is incorrect. (3) The model trained on Russian targets can use English data to form sentences in Russian, transcribing the entities to Cyrillic.

		BLEU		**METEOR**		**ChrF++**		**TER**		**BERTScore**		**BLEURT**	
All	Ours	50.34	(10)	0.398	(8)	0.666	(8)	0.435	(7)	0.951	(8)	0.57	(8)
	Baseline	40.57	(14)	0.373	(15)	0.621	(15)	0.517	(14)	0.943	(14)	0.47	(12)
Seen Cat.	Ours	59.13	(10)	0.422	(10)	0.712	(9)	0.403	(7)	0.960	(9)	0.58	(14)
	Baseline	42.95	(31)	0.387	(27)	0.650	(28)	0.563	(31)	0.943	(31)	0.41	(31)
Unseen Cat.	Ours	42.24	(10)	0.375	(13)	0.617	(10)	0.46	(7)	0.943	(11)	0.52	(10)
	Baseline	37.56	(12)	0.357	(15)	0.584	(15)	0.51	(13)	0.940	(12)	0.44	(12)
Unseen Ent.	Ours	51.23	(4)	0.406	(8)	0.687	(7)	0.417	(9)	0.959	(8)	0.63	(8)
	Baseline	40.22	(17)	0.384	(15)	0.648	(15)	0.476	(14)	0.949	(13)	0.55	(12)

Table 2: Results of our approach on English (all data, seen categories, unseen categories, unseen entities), compared to the baseline. The numbers in brackets show the rank of each model (out of 35 submissions) with respect to the given metric.

Similarly to Freitag and Roy (2018), we observe that in English, linearized triples can be seen as a noisy version of the output text, where:

- subjects and objects are copied verbatim,
- predicates are shortened or reworded,
- function words are deleted,
- order of the entities is shuffled.

mBART's pretraining objective is different from this, but we hypothesize that it is similar enough to be relevant for our task. For denoising Russian, our intuition stems from mBART's successful application in machine translation (Liu et al., 2020).

We finetune the pre-trained `mbart.CC25`[4] model from the FAIRSEQ toolkit (Ott et al., 2019). We follow the example instructions for finetuning the model, changing only the `total_updates` to 10,000 to reflect the smaller size of our data. We show the capabilities of our model in Table 1.

5 Results

We report on WebNLG automatic and human evaluation results, as well as our own error analysis.

5.1 Automatic Metrics

Automatic metrics used in the challenge include BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ChrF++ (Popović, 2017), TER (Snover et al., 2006), BERTScore (Zhang et al., 2020), and BLEURT (only used for English; Sellam et al., 2020). The results of our approach for English are shown in Table 2, comparing to the baseline.[5] We can see that our approach comfortably beats the baseline in all metrics and places in the first third of the submissions. While it does lose performance on unseen categories, the drop is not as dramatic as for many other competing approaches; our system is able to hold or improve its rank in the results table. Compare the baseline's ranking for seen categories, where it placed near the bottom of the list, and the ranking for unseen categories, where it scores in the first half – this shows that many approaches fared worse than the baseline on unseen categories, unlike our system.

The results for Russian are shown in Table 3. There were fewer submissions for Russian, and our system not only beats the baseline by a large

[4] `https://github.com/pytorch/fairseq/tree/master/examples/mbart`

[5] See `https://gerbil-nlg.dice-research.org/gerbil/webnlg2020results` for full results.

	BLEU		**METEOR**		**ChrF++**		**TER**		**BERTScore**	
Ours	52.93	(1)	0.672	(2)	0.677	(2)	0.398	(1)	0.909	(1)
Baseline	23.53	(12)	0.461	(12)	0.511	(12)	0.680	(12)	0.836	(12)

Table 3: Results of our approach on Russian data, compared to the baseline. The numbers in brackets show the rank of each model (out of 12 submissions) if ordered by the given metric.

margin (as did all competing submissions), but it is able to rank first in 3 metrics out of 5 (BLEU, TER, BERTScore) and second in the remaining ones.

5.2 Human Evaluation

The challenge organizers ran a human evaluation campaign[6], where annotators were asked to rate-five aspects of the output texts: data coverage, relevance, correctness, text structure and fluency. Each criterion has been rated with a number in the range from "0" (completely disagree) to "100" (completely agree). The scores were clustered into groups (1-5; 1 being the best) among which there are no statistically significant differences according to the Wilcoxon rank-sum test (Wilcoxon, 1992).

Our systems placed in the top clusters (1 or 2) for both English and Russian. For English, our system ranks first for all the categories in *seen domains*, and first or second in *unseen entities* and *unseen domains*. In total, our English system achieved rank 1 for relevance, correctness and text structure, and rank 2 for data coverage and fluency. For Russian, our system ranks second for correctness and first in all other categories.

5.3 Manual Analysis

To better understand the nature of errors made by our system, we manually inspected a sample of 50 outputs in each language.[7] We found factual errors in 12 English outputs, mostly concentrated along the unseen categories (*Scientist*, *Movie*, *Musical Record*). The model tends to describe musical works and movies in terms of written works ("written", "published" etc.), i.e., the closest seen category. There are also several swaps in roles of the entities (e.g., "is to southeast" instead of "has to its southeast", "follows" instead of "is followed by" etc.). In a few cases, the model hallucinates a relation not specified in the data (e.g., "born on January 1, 1934 in Istanbul" when a date of birth and current residence is given, not the birthplace) or is not able to infer background knowledge not given on the input (it talks about a dead person in the present tense). The swaps in roles and hallucinated relations also occured in Russian; in addition, we found a hallucinated (correct) airport name and a few forgotten ingredients for a dish from a long list. Factual errors in Russian were less frequent (9 sentences), which is expected as there are no unseen categories. Moreover, the system shows an impressive performance at translating entity names from the English RDF into Russian.

We further found 10 outputs with suboptimal phrasing in English and 9 in Russian, where the model did not connect properties of the same type in a coordination (e.g., two musical genres for a record) or gave numbers without proper units (e.g., "runtime of 89.0" or "area of 250493000000.0").

6 Discussion

Our solution benefits from the denoising skills of the pre-trained mBART model, which to a certain extent combines all the tasks of the micro-planning pipeline (lexicalization, aggregation, surface realization, referring expression generation, sentence segmentation). Finetuning on task-specific data then mostly helps to specify the task at hand. Moreover, multilingual pre-training allows us to use a single architecture for both English and Russian.

That being said, we note the RDF-to-text task is far from solved. The performance of our model is noticeably lower on categories unseen in training, and it is prone to swapping relations of entities or hallucinating relations. Even though the longest examples in the WebNLG dataset fit into the model, the length of the input sequence is still limited and the model does not generalize for inputs of arbitrary size. Moreover, English and Russian are coincidentally the two most represented languages in the mBART pre-training corpora (ca. 300 GB of data each) and the performance of our model would probably be lower with low-resource languages.

[6] See `https://beng.dice-research.org/gerbil/webnlg2020resultshumaneval` for full results.

[7] While one of the authors has some knowledge of Russian, it is nowhere near a native level. Automatic back-translation to English was used in a few cases to facilitate understanding.

7 Conclusion

We presented a simple setup for RDF-to-text generation, consisting of triple linearization and text denosing. With the help of a multilingual pretrained model, this approach is language-agnostic and yields high-quality results with minimal effort. We hope that it will serve as a baseline for more complex approaches to RDF-to-text generation.

Acknowledgements

This work was supported by the Charles University GAUK grant No. 140320, the SVV project No. 260575, and the Charles University project PRIMUS/19/SCI/10.

References

Thiago Castro-Ferreira, Claire Gardent, Nikolai 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)*, Online.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

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.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

Markus Freitag and Scott Roy. 2018. Unsupervised Natural Language Generation with Denoising Autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3922–3929, Brussels, Belgium.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for NLG micro-planning. In *55th annual meeting of the Association for Computational Linguistics (ACL)*, pages 179–188, Vancouver, Canada.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. 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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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 Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online.

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. *arXiv preprint arXiv:2001.08210*.

Diego Moussalem, Paramjot Kaur, Thiago Castro-Ferreira, Chris van der Lee, Anastasia Shimorina, Felix Conrads, 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)*, Online.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and

Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, MN, USA.

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, PA, USA.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Thibault Sellam, Dipanjan Das, and Ankur P. 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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA, USA.

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 (NeurIPS)*, pages 5998–6008, Long Beach, CA, USA.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12).

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. CC-Net: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 4003–4012, Marseille, France.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Online.