# Better Translation + Split and Generate for Multilingual RDF-to-Text

**Nalin Kumar**[♣], **Saad Obaid ul Islam**[♠], **Ondřej Dušek**[♣]
[♣]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
[♠]Center For Artificial Intelligence and Data Science, University of Würzburg, Germany
kumarnal@student.cuni.cz, saad.obaid-ul-islam@uni-wuerzburg.de, odusek@ufal.mff.cuni.cz

## Abstract

This paper presents system descriptions of our submitted outputs for WebNLG Challenge 2023. We use mT5 in multi-task and multilingual settings to generate more fluent and reliable verbalizations of the given RDF triples. Furthermore, we introduce a partial decoding technique to produce more elaborate yet simplified outputs. Additionally, we demonstrate the significance of employing better translation systems in creating training data.

## 1 Introduction

State-of-the-art in natural language generation (NLG) recently has benefited from self-supervised pretrained language models (PLMs) such as BART, T5, and GPT-2 (Lewis et al., 2020; Raffel et al., 2020; Radford et al., 2019), which provided performance improvements on many NLG tasks (Zhang et al., 2020b; Wei et al., 2021; Kale and Rastogi, 2020). One of the most prominent NLG tasks is data-to-text generation (Wiseman et al., 2017), with multiple public datasets, including WebNLG (Colin et al., 2016), ToTTo (Parikh et al., 2020), or DART (Nan et al., 2021). A significant problem of these datasets is their limitation to English only, which effectively rules out their application to other languages; incorporating low-resource languages in NLG is generally an open problem (Ruder, 2019).

Researchers addressed this by multilingual PLMs like mT5 (Xue et al., 2021) and mBART (Liu et al., 2020) pre-trained on massively multilingual corpora like Common Crawl 101.[1] Unfortunately, given the disproportionate representation of low-resource languages in multilingual datasets,[2] the efficacy of PLMs on these languages is compromised. Efforts are being made to improve PLM performance for low-resource languages via cross-lingual transfer (Conneau et al., 2020). To assess the state-of-the-art performance and spur development in data-to-text NLG techniques for low-resource languages, the WebNLG 2023 challenge features five non-English languages – four entirely new low-resource ones: Maltese (*Mt*), Welsh (*Cy*), Irish (*Ga*) and Breton (*Br*), with the addition of Russian (*Ru*) from the 2020 iteration (Castro Ferreira et al., 2020). While development and test sets for the low-resource languages were created by hand, training sets are machine-translated.

In our submission of WebNLG 2023 challenge, we use the power of mT5 (Xue et al., 2021), a multilingual transformer model trained in a multi-task fashion. To boost performance on low-resource languages, we employ several additional steps in fine-tuning and inference: (1) We improve WebNLG 2023 training data by re-translating with a stronger machine translation (MT) model (Costa-jussà et al., 2022). (2) We finetune either individual models for each language (single-task setting), or a single model for all languages (multi-task setting). (3) We generate text by splitting the input RDF triples and decoding a subset of triples at a time. We publish our code and submitted outputs on GitHub.[3]

## 2 Related Works

A lot of non-English NLG works addresses either surface realization (Mille et al., 2018; Fan and Gardent, 2020) or text-to-text tasks such as summarization (Straka et al., 2018; Scialom et al., 2020; Hasan et al., 2021) or question generation (Shakeri et al., 2021). In data-to-text generation, Dušek and Jurčíček (2019) trained a RNN-based model for Czech, van der Lee et al. (2020) apply a similar architecture for Dutch. A RNN decoder has also been applied to Japanese in a multimodal setup (Ishigaki et al., 2021). Recent works mostly

[1] https://commoncrawl.org/
[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages
[3] https://github.com/knalin55/CUNI_Wue-WebNLG23_Submission



*Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 73–79
September 12, 2023 

use transformer-based architectures. Kale and Roy (2020) applied a custom transformer pretrained for MT to a Czech data-to-text task. Several systems targeted WebNLG 2020's Russian task (Castro Ferreira et al., 2020). Agarwal et al. (2020)'s system features a custom-trained bilingual T5-derived transformer, the works of Kasner and Dušek (2020) and Zhou and Lampouras (2021) build on the standard mBART model. Our approach combines the use of MT for preprocessing with standard multilingual PLMs and adds multi-task learning and sentence-level generation.

## 3 WebNLG 2023 Task Description

The WebNLG challenge focuses on generating text from a set of RDF triples. Input RDF triples are extracted from DBpedia (Auer et al., 2007) and the corresponding reference texts are gathered through crowdsourcing. Previous iterations (Gardent et al., 2017; Castro Ferreira et al., 2020) focused on English and Russian. The 2023 WebNLG challenge includes 4 low-resource languages (*Mt*, *Cy*, *Ga*, *Br*), along with Russian (*Ru*). The development and test sets are manually translated from English, but the training data is a result of MT by the Edinburgh Zero translation system (Zhang et al., 2020a).

### 3.1 Automatic Metrics

WebNLG 2023 has been evaluated by automatic metrics so far. These include BLEU (Post, 2018), METEOR (Lavie and Agarwal, 2007), chrF++ (Popović, 2015), and TER (Snover et al., 2006). Comparison against multiple references is used for *Ru*, other languages use a single reference per instance.

### 3.2 Organizer Baselines

The organizer-provided baselines for *Ru* include Kasner and Dušek (2020)'s finetuned mBART and the 2020 Challenge baseline, the FORGE system (Mille et al., 2019) coupled with Google Translate. The baseline for *Mt*, *Ga*, *Cy*, *Br* is the 2020 Challenge English system of Guo et al. (2020), coupled with Edinburgh Zero MT (Zhang et al., 2020a).

## 4 Proposed Approaches

Our own approach builds on multilingual transformer PLMs, but improves input data processing in the low-resource languages by better translation or filtering (Section 4.1) and employs two simple yet effective strategies to improve generation: multitask learning (Section 4.2) and splitting complex inputs for decoding (Section 4.3).

### 4.1 Better Translation & Data Filtering

To improve on the challenge's generate-and-translate baseline (see Section 3.2) and show the effect of the quality of MT-processed training data on the outputs in low-resource languages, we replace or filter the baseline MT system outputs. We then use these improved MT outputs both in an alternative generate-and-translate baseline and as improved training data for our direct NLG methods.

Specifically, for *Mt*, *Ga*, and *Cy*, we replace the Edinburgh Zero System (Zhang et al., 2020a) with the state-of-the-art NLLB system (Costa-jussà et al., 2022). Additionally, we modify the translation process. The original training data was created by translating *whole En verbalizations*. However, based on our inspection, the resulting translations are often incomplete. We counter this problem by translating *individual En sentences* using NLLB.

For *Br* where NLLB is unavailable, we filter out inconsistent examples from the existing MT-processed training data. Since each input set of triples in the training set typically corresponds to multiple verbalizations, we calculate the ratios of lengths of all *En* verbalizations to their corresponding *Br* translations. Subsequently, we filter out the *Br* verbalizations with ratios (to their corresponding *En* original) smaller than half the maximum ratio for the given input triple set.

### 4.2 Multitask Learning (MTL)

Multitask learning (MTL) trains models on diverse tasks simultaneously, improving generalization to different tasks and domains (Liu et al., 2019; Raffel et al., 2020; Sanh et al., 2022). We apply MTL to improve the model's understanding of input triples: In addition to data-to-text generation in the target language, we use translation from English and data-to-text in English as auxiliary tasks. The model learns to distinguish tasks by different prompts.

### 4.3 Split and Generate (SaG)

Due to data quality problems being more common in larger input triple sets, PLMs struggle to generate fluent and accurate outputs for complex inputs. In a previous study, Narayan et al. (2017) introduced a split and rephrase approach for sentence simplification. Their method involved creating parallel data by finding the verbalizations of a subset of input

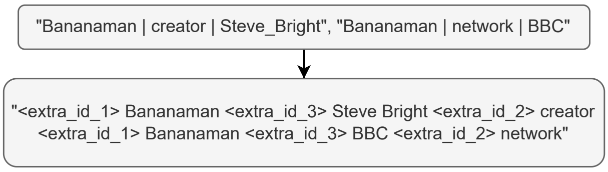


Figure 1: An instance of the preprocessing step

triples in the WebNLG dataset. Building upon this concept and following earlier works on sentence-by-sentence verbalization (Moryossef et al., 2019; Ferreira et al., 2019), we employ a straightforward decoding strategy: We split the input triple sets into subsets based on the same triple subject, generate outputs for the individual subsets, and subsequently concatenate the generated texts. We further experiment with a subset size limit – we partition the subset $S$ into further subsets if its size exceeds a pre-set value $n$.

## 5 Experiments

### 5.1 Data preprocessing

As is common in PLM-based RDF-to-text systems, we linearize the input triples. We use three special tokens (newly added to the model) to mark subjects, objects, and predicates. We remove underscores from entities. The triples are concatenated using a subject-object-verb order (see Figure 1). Following mT5 pretraining methodology, we use prompts to distinguish tasks. For RDF triple verbalization, we adopt the prompt format: "*RDF-to-text in <lang>: <input>*" where *<lang>* is the target language and *<input>* denotes the input triples.

### 5.2 Model Parameters

We use the base variant of mT5 (Xue et al., 2021).[4] We train all our models for 30 epochs with batch sizes of 8 and select our best model based on validation loss. We use beam search decoding with beam size 4. To prevent repetitive outputs, we also use a repetition penalty $rp = 3.5$ (Keskar et al., 2019).[5]

### 5.3 Model Variants

We experiment with the following system variants (in both training and inference):

**baseline** For each language, we simply fine-tune mT5 on the organizer-provided training data.

**base-(NLLB/DF)** For *Mt*, *Ga* and *Cy*, we retranslate the training dat using NLLB (Costa-jussà et al., 2022) and for *Br*, we apply simple data filtering, as described in Section 4.1. We do not apply this step for *Ru* where the training data is of sufficient quality. Since the performance of retranslated and filtered data is better than the baseline (see Tables 1b to 1d, Table 2), we use it as our primary training data in all further experiments.

**MTL** We finetune separate mT5 models using MTL for *Mt*, *Ga*, and *Cy* (using NLLB-retranslated data) and *Ru* (on original data), with translation to English and English RDF-to-text as auxiliary tasks (see Section 4.2). For translation from English, we use the prompt "*Translate from En to <lang>: <input>*", and the prompt for English text generation is "*RDF-to-text in English: <input>*".

**Multilingual Model (MLM)** We finetune a single mT5 model on data created from combining NLLB-retranslated *Mt*, *Ga*, and *Cy* datasets with the existing *Ru* data. Since the *Br* data is lower quality, we do not include it in this setting.

**SaG** We optionally employ SaG during inference to ensure simpler and more fluent outputs (see Section 4.3). We only use the subset size limit $n$ in *Br*, as we observed no performance improvements in other languages. For *Br*, we find that setting $n = 2$ yields the best performance.

## 6 Results

Table 1 shows comprehensive evaluation scores of our experiments on *Ru*, *Mt*, *Ga* and *Cy*. Generally, we see a substantial improvement for *base-NLLB* from *baseline* for *Mt*, *Ga*, and *Cy*. This is the most notable boost as using NLLB represents the largest change in the setup.

For *Ru* (Table 1a), we observe an improvement with the *MTL* and *MLM* setups, but *SaG* decoding seems to hurt performance. As *Ru* is a comparatively high-resource language, we suspect the model is already well-suited to process complex outputs in a single step, and splitting the inputs means losing context in the individual steps.

Unlike *Ru*, *Mt* (Table 1b) sees a performance drop of *MTL/MLM* over *base-NLLB*. Adding SaG helps slightly, but not enough to match *base-NLLB*. The organizers' baseline and generate-and-translate with NLLB are better than any of our models.

For both *Ga* and *Cy*, we observe an improvement in the performance of *MTL* and *MTM* over *base-*

[4] https://huggingface.co/google/mt5-base
[5] The value was found by development data trials.

| Model | BLEU | MET | ChrF | TER |
|---|---|---|---|---|
| WebNLG 2023 best | 54.71 | - | 0.69 | 0.37 |
| (Kasner and Dušek, 2020) | 52.9 | - | 0.68 | 0.40 |
| (Mille et al., 2019) + GT | 25.5 | - | 0.51 | 0.67 |
| baseline | 51.39 | 0.37 | 0.67 | 0.41 |
| MTL | 53.48 | 0.39 | 0.68 | 0.4 |
| MLM ◇ | **54.52** | **0.39** | **0.69** | **0.38** |
| MTL + SaG | 50.15 | 0.38 | 0.67 | 0.44 |
| MLM + SaG | 50.12 | 0.38 | 0.68 | 0.44 |

(a) Russian (*Ru*) – 2nd of 4 submitted systems

| Model | BLEU | MET | ChrF | TER |
|---|---|---|---|---|
| WebNLG 2023 best | 21.27 | - | 0.52 | 0.65 |
| (Guo et al., 2020) + Edin | 15.60 | - | 0.42 | **0.67** |
| (Guo et al., 2020) + NLLB | **16.07** | 0.26 | **0.47** | 0.71 |
| baseline | 12.37 | 0.2 | 0.36 | 0.72 |
| base-NLLB | 14.08 | 0.25 | 0.44 | 0.77 |
| MTL | 13.95 | 0.25 | 0.44 | 0.78 |
| MLM | 13.91 | 0.25 | 0.45 | 0.77 |
| MTL + SaG ◇ | 14.02 | 0.26 | 0.45 | 0.78 |
| MLM + SaG | 14.02 | **0.29** | 0.45 | 0.79 |

(b) Maltese (*Mt*) – 3rd of 5 submitted systems

| Model | BLEU | MET | ChrF | TER |
|---|---|---|---|---|
| WebNLG 2023 best | 20.40 | - | 0.51 | 0.69 |
| (Guo et al., 2020) + Edin | 11.63 | - | 0.36 | 0.74 |
| (Guo et al., 2020) + NLLB | **17.95** | 0.23 | **0.46** | **0.70** |
| baseline | 6.53 | 0.13 | 0.27 | 0.77 |
| base-NLLB | 15.65 | 0.22 | 0.43 | 0.78 |
| MTL | 15.65 | 0.22 | 0.43 | 0.77 |
| MLM | 15.70 | 0.22 | 0.43 | 0.78 |
| MTL + SaG ◇ | 15.87 | 0.22 | 0.43 | 0.78 |
| MLM + SaG | 15.15 | **0.32** | 0.42 | 0.8 |

(c) Irish (*Ga*) – 3rd/4th of 5 submitted systems

| Model | BLEU | MET | ChrF | TER |
|---|---|---|---|---|
| WebNLG 2023 best | 25.11 | - | 0.55 | 0.64 |
| (Guo et al., 2020) + Edin | 10.70 | - | 0.36 | 0.77 |
| (Guo et al., 2020) + NLLB | **18.77** | 0.25 | **0.48** | **0.7** |
| baseline | 7.80 | 0.15 | 0.29 | 0.78 |
| base-NLLB | 16.13 | 0.24 | 0.44 | 0.79 |
| MTL | 16.73 | 0.24 | 0.45 | 0.78 |
| MLM | 16.65 | 0.24 | 0.44 | 0.80 |
| MTL + SaG ◇ | 17.01 | 0.24 | 0.45 | 0.79 |
| MLM + SaG | 16.28 | **0.35** | 0.45 | 0.81 |

(d) Welsh (*Cy*) – 3rd of 4 submitted systems

Table 1: Automatic Evaluation Scores for *Ru*, *Mt*, *Ga*, *Cy*. "MET" stands for METEOR. We include the scores for the best WebNLG 2023 system and baseline systems above the dividing line in each table. For *Ru*, these come from WebNLG 2020. The English outputs of (Mille et al., 2019)'s system were processed by Google Translate (GT). For *Mt*, *Ga*, *Cy*, the English outputs of (Guo et al., 2020)'s system were processed by Edinburgh Zero (Zhang et al., 2020a) (Edin) and re-processed by ourselves using NLLB (Costa-jussà et al., 2022), see Section 4.1. ◇ denotes our system used for the challenge submission. The rankings of our submission and the number of submitted systems are included in each sub-caption.

| Model | BLEU | MET | ChrF | TER |
|---|---|---|---|---|
| (Guo et al., 2020) + Edin | 9.92 | - | 0.32 | **0.77** |
| baseline | 7.02 | 0.14 | 0.27 | 0.79 |
| base-DF | 8.84 | 0.15 | 0.29 | 0.91 |
| base-DF + SaG ◇ | 10.09 | 0.17 | 0.33 | 0.80 |
| base-DF + SaG ($n = 2$) | **11.31** | **0.19** | **0.36** | 0.83 |

Table 2: Automatic evaluation scores for *Br* (see Table 1 for explanations). Our submission was the only one competing in the challenge.

*NLLB* (Tables 1c and 1d), similar to *Ru*. There is a slight improvement in the performance of *SaG* decoding over *MTL*, but not in the *MTM* case. While our direct translation models outperform the challenge baseline, they fare worse than generate-and-translate with NLLB.

The scores for *Br* are shown in Table 2. Using *SaG* decoding with base-DF slightly enhances the performance over the organizer's baseline, and unlike other languages, *Br* sees a decent performance boost over the method without *SaG* decoding.

## 7 Conclusion

In this work, we describe our submitted systems for the WebNLG 2023 Challenge. The provided silver training dataset for the under-resourced languages sometimes contains incomplete translations. To address this, we create alternate parallel data using NLLB, which seems to provide the biggest performance boost. Furthermore, we gain additional minor improvements by using a multi-task and multilingual training settings, coupled with a split-and-generate decoding method, which produces simpler and more verbose outputs. Our systems mostly score in the middle of the challenge scoreboard; our *Br* submission was the only competitor in the challenge. For *Ru*, *Ga* and *Cy*, our submitted systems perform substantially better than the organizers' baseline. A non-submitted variant of our system for *Br* also surpasses the baseline. However, in *Mt*, *Ga*, and *Cy*, our approaches underperform a simple generate-and-translate baseline with the improved NLLB MT system.

## Acknowledgements

This work was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG). It used resources provided by the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth and Sports project No. LM2018101).

## 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, Ireland (Virtual). Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

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.

Emilie Colin, Claire Gardent, Yassine M'rabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The WebNLG challenge: Generating text from DBPedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, Edinburgh, UK. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ondřej Dušek and Filip Jurčíček. 2019. Neural Generation for Czech: Data and Baselines. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG 2019)*, pages 563–574, Tokyo, Japan. ArXiv: 1910.05298.

Angela Fan and Claire Gardent. 2020. Multilingual AMR-to-Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online. 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 *2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) and 9th International Joint Conference on Natural Language Processing (IJCNLP)*, Hong Kong. ArXiv: 1908.09022.

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.

Qipeng Guo, Zhijing Jin, Ning Dai, Xipeng Qiu, Xiangyang Xue, David Wipf, and Zheng Zhang. 2020. P2: A Plan-and-Pretrain Approach for Knowledge Graph-to-Text Generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 100–106, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating Racing Game Commentary from Vision, Language, and Structured Data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Mihir Kale and Scott Roy. 2020. Machine Translation Pre-training for Data-to-Text Generation – A Case Study in Czech. In *Proceedings of the 13th International Conference on Natural Language Generation*, Online. ArXiv: 2004.02077.

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, Online.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

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. 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.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

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.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The First Multilingual Surface Realisation Shared Task (SR'18): Overview and Evaluation Results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia.

Simon Mille, Stamatia Dasiopoulou, Beatriz Fisas, and Leo Wanner. 2019. Teaching FORGe to Verbalize DBpedia Properties in Spanish. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 473–483, Tokyo, Japan. Association for Computational Linguistics.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-Step: Separating Planning from Realization in Neural Data-to-Text Generation. In *NAACL*, Minneapolis, MN, USA. ArXiv: 1904.03396.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Sebastian Ruder. 2019. The 4 biggest open problems in nlp. `https://www.ruder.io/4-biggest-open-problems-in-nlp/`.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The Multilingual Summarization Corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Siamak Shakeri, Noah Constant, Mihir Kale, and Linting Xue. 2021. Towards Zero-Shot Multilingual Synthetic Question and Answer Generation for Cross-Lingual Reading Comprehension. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 35–45, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich 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: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtský, Vojtěch Hudeček, and Jan Hajič. 2018. SumeCzech: Large Czech News-Based Summarization Dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chris van der Lee, Chris Emmery, Sander Wubben, and Emiel Krahmer. 2020. The CACAPO Dataset: A Multilingual, Multi-Domain Dataset for Neural Pipeline and End-to-End Data-to-Text Generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 68–79, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020b. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Giulio Zhou and Gerasimos Lampouras. 2021. Generalising Multilingual Concept-to-Text NLG with Language Agnostic Delexicalisation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 114–127, Online. Association for Computational Linguistics.