Evaluating Knowledge Graph-to-text Generation Models for English and Russian on Out Of Domain Data

Anna Nikiforovskaya
Loria, CNRS; Université de Lorraine
Nancy, France
anna.nikiforovskaja@loria.fr

Claire Gardent
Loria, CNRS
Nancy, France
claire.gardent@loria.fr

Abstract

While the WebNLG dataset has prompted much research on generation from knowledge graphs, little work has examined how well models trained on the WebNLG data generalise to unseen data and work has mostly been focused on English. In this paper, we introduce novel benchmarks for both English and Russian which contain various ratios of unseen entities and properties. These benchmarks also differ from WebNLG in that some of the graphs stem from Wikidata rather than DBpedia. Evaluating various models for English and Russian on these benchmarks shows a strong decrease in performance while a qualitative analysis highlights the various types of errors induced by non i.i.d data.

1 Introduction

Knowledge graphs (KGs) describe connections between entities (e.g., people, places or events) thereby representing knowledge about the world. The task of KG-to-Text generation consists in verbalising the content of a KG. Much research on KG-to-Text generation focuses on the WebNLG dataset (Gardent et al., 2017) often restricting evaluation to the WebNLG test sets. While these include both seen (in domain) and unseen (out of domain, OOD) data for English, no unseen test data is available for Russian. Furthermore, the input graphs all stem from DBpedia and the texts are often stilted as they are either crowd-sourced (English data) or machine translated from the crowdsourced texts and manually verified (Russian data).

To assess how well current NLG models perform on OOD KG-to-Text generation, we create several novel benchmarks for both English and Russian which address these shortcomings and differ from the WebNLG test sets in several ways. First, they include both English and Russian – WebNLG only has unseen test data for English. Second, they include both DBpedia and Wikidata[1] graphs – WebNLG focuses on DBpedia graphs. Third, they contain various ratios of unseen entities and properties – this allows for a detailed analysis of how the type and ratio of unseen data impact performance.

Using these benchmarks, we then assess and compare several KG-to-Text models. The results show a strong decrease in performance for all models compared to results on in domain data. A qualitative analysis highlights the various types of errors induced by OOD data suggesting directions for further research on KG-to-Text.

2 Related Work

KG-to-Text Generation. The WebNLG challenges gave rise to different approaches for KG-to-Text generation, most of the 2020 participating models being fine-tuned version of T5 (Raffel et al., 2020) or BART (Lewis et al., 2020). In the WebNLG 2020 challenge (Castro Ferreira et al., 2020), human evaluation showed that models which were based on these pre-trained encoder-decoders produce the best texts in terms of fluency (e.g., Yang et al. (2020); Agarwal et al. (2020)) but lacked adequacy on unseen test sets exposing a noticeable drop in performance regarding Relevance (not all information mentioned in the text is present in the input graph) and Data Coverage (not all information present in the input graph is verbalised by the text).

For Russian, the two best performing models are Kazakov et al. (2023) and Kumar et al. (2023). Both models fine-tune a pre-trained model on the WebNLG data with Kazakov et al. (2023) fine-tuning the pre-trained FRED (Full-scale Russian Enhanced Denoiser, 1.7M Parameter) model and Kumar et al. (2023) mT5_{base}. Neither of these models were evaluated on unseen data.

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

Proceedings of the 17th International Natural Language Generation Conference, pages 134–144
September 23–27, 2024.

Evaluation. Recent work has focused on creating better evaluation benchmarks for data-to-text generation. In particular, Mille et al. (2021) introduced various subtests (subpopulations) for different data-to-text generation tasks including WebNLG. They developed subpopulations based on input size and the uniqueness of subjects, objects, and properties present in the data. Their study showed that each of these properties influences the results and that the level of impact differs between Russian and English. Similarly, in 2024, a new GEM challenge on Data-to-Text generation was launched which includes parallel datasets to WebNLG featuring counterfactual and fictional data.[2] This challenge also evaluates data-to-text generation models on graphs from Wikidata (Axelsson and Skantze, 2023). These new test sets consist solely of automatically combined graphs without any reference verbalizations, which excludes reference-based evaluation and necessitate human evaluation.

Different from these works, we provide new unseen test sets for KG-to-Text generation which include references in both English and Russian. We then used these test sets to evaluate the ability of existing models to generalise to OOD data and to analyse the types of errors that arise in their output texts.

3 Creating New Benchmarks for English and Russian

We aim to create benchmarks which support a fine-grained assessment of how various types of unseen items impact generation.

Terminology. An *unseen element* is a KG element (entity or property) not seen in the WebNLG training/dev data. An *unseen category* is a DBpedia category which is not part of the 16 categories[3] used in WebNLG to create the training data.

We create separate benchmarks depending on whether the input graph contains unseen entities, unseen entities and properties or unseen category. While the latter two benchmarks permit assessing how well models perform on out of domain data, the former helps evaluating how much performance degrades with varying ratios of unseen entities.

For English and Russian, we derive these benchmarks from the KELM dataset (Agarwal et al., 2021), a large dataset of (graph,text) pairs created using distant supervision. For Russian, we additionally derive benchmarks from the WebNLG data following a methodology similar to that used to create the WebNLG unseen test set for English.

KELM. Agarwal et al. (2021) created the KELM dataset in several steps as follows. First, Wikidata triples were heuristically aligned to Wikipedia sentences yielding a dataset of approximately 6M noisily aligned (graph, sentence) pairs and covering 1,041 Wikidata properties. Second, 15M Wikidata graphs where created based on relation co-occurrence counts and the corresponding text was generated from these graphs using a T5 model fine-tuned on the silver 6M (graph,sentence) pairs. The semantic adequacy (semantic match between graph and text) and the fluency of 200 randomly selected KELM (graph,text) pairs were annotated by human judges (8 annotators, 2 judgements per instance) on a 1-5 scale, yielding an average rate of 4.36 for semantic adequacy and 4.60 for fluency. Examples of KELM instances are shown in table 1.

WebNLG. The WebNLG dataset is a dataset of (graph,text) pairs where graphs were extracted from DBpedia and texts were crowdsourced to match the input graph. For English, the training data covers 16 DBpedia categories and the test set has three subsets: Seen (490 instances), a test set where graphs include only entities and properties present in the training data; Unseen Entities (393 instances), where graphs include entities not present in the training data; and Unseen Categories (896 instances), a test set where graphs are rooted in entities whose category does not belong to the 16 categories present in the training data.[4] For Russian, the training data only covers nine categories[5] and all instances in the test set (1,200 instances) are from the seen categories.

4 Creating Kelm Benchmarks

To create the KELM unseen test sets (KELM-E, KELM-E+P), we first select subsets of KELM that contain unseen entities and properties. We then ask human annotators to verify the semantic adequacy of the (graph, text) pairs (does the text match

[2] https://gem-benchmark.com/shared_task

[3] The 16 categories used to anchor WebNLG data are: Airport, Astronaut, Building, City, ComicsCharacter, Food, Monument, SportsTeam, University, WrittenWork, Athlete, Artist, CelestialBody, MeanOfTransportation, Politician, Company.

[4] For each test set there are two versions, one for generation and the other for semantic parsing. Here we only consider the generation test sets.

[5] These nine DBpedia categories are: Airport, Astronaut, Building, CelestialBody, ComicsCharacter, Food, Monument, SportsTeam, and University.

Text	Graph
The redshift of NGC 266 is 0.015537.	(*NGC 266* , *redshift*, *0.015537*)
Bowditch is a lunar crater which is located at LQ22 on the Moon and named after Nathaniel Bowditch.	(*Bowditch_crater*, *located on astronomical location*, *Moon*), (*Bowditch_crater*, *instance of*, *Lunar craters*), (*Bowditch_crater*, *location*, *LQ22*), (*Bowditch_crater*, *named after*, *Nathaniel Bowditch*)

Table 1: Examples from KELM dataset

the graph?) filtering out all pairs which are not validated by the annotators. This yields novel unseen test sets for English. We create corresponding test sets for Russian using machine translation and manual correction by professional translators.[6]

In what follows, KELM refers to the dataset created by (Agarwal et al., 2022) while KELM-E, KELM-E+P refers to the two benchmarks we derived from KELM.

Selecting a Subset of KELM. We extract a subset of KELM such that (i) graph and text embeddings have high similarity, (ii) the dataset is balanced across graph size and (iii) the distribution of the Wikidata properties present in the KELM dataset is preserved. The latter point helps ensuring that our dataset has a wide variety of topics and is not skewed towards frequent properties.

To extract this subset, we proceed as follows. First, we compute graph and text embeddings using Le Scao and Gardent (2023) cross-modal KG-Text model and we only keep those pairs whose graph and text embeddings have a cosine similarity greater than 0.9. We then remove quadruples (i.e., Wikidata facts that are not triples) and graphs that have more than six triples[7] as these are a minority (less than 1%) and tend to have repetitive or unintelligible texts. We further compute the ratio of unseen elements for each graph text pairs. Finally, we select two types of unseen data: instances where all properties are known but some entities are not (unseen entities, KELM-E) and instances which contain various ratio of unseen entities and properties (unseen entities and properties, KELM-E+P).

[6]An alternative would be to create a Russian dataset from Wikidata and Wikipedia using (Agarwal et al., 2022) methodology. We adopted the MT approach instead because it is less computationally intensive and it allows for the creation of a parallel (graph, English text, Russian text) dataset.

[7]Creating a dataset for larger graphs is possible but would require developing an alternative content selection procedure to ensure that the selected subgraphs yield text that are coherent and readable.

Human Validation on English Data. A manual inspection of 100 random instances shows that approximately one third of the data is poorly aligned i.e., text and graph convey different content. We use crowd sourcing to filter out badly aligned (graph,text) pairs. We use the Potato annotation tool (Pei et al., 2022) to create a website for annotation and Prolific[8] to find participants for the study. We provide a screenshot of the built website in Appendix A. The participants were paid 14€ for annotating 100 instances and 2€ for the qualification task (even if failed) which averages to 10.5€ per hour. Further details about the human annotation protocol are given in Section A.

To evaluate the quality of each pair, we used the WebNLG Challenge 2023 criteria for human evaluation (Cripwell et al., 2023) whereby for each item, the annotators were asked to answer the following four questions (with binary yes/no answer for the first three questions).

No omission. "Looking at each element of the graph in turn, does the text express each of these elements in full (allow synonyms and aggregation)?".

No addition. "Looking at the text, is all of its content expressed in the graph? (Allow duplication of content.)".

No unnecessary repetition. "Is any content in the text unnecessarily repeated?".

Fluency. " Please rate the text shown in terms of fluency on a scale of 1 to 5 where 5 is the highest (best) score. Highly fluent text 'flows well' and is well connected and free from disfluencies.".

To ensure a good understanding of these criteria, we made available an annotation codebook with explanation and examples for each criterion. We also run a prestudy consisting of 15 (graph,text) pairs where 10 examples were taken from KELM and 5 easier examples were created manually. We made sure that the examples covered all possible

[8]https://prolific.com/

answers for each yes/no criteria and were of different level of fluency. However, as it is hard to evaluate fluency, we only verified if the participant answered to all yes/no criteria correctly. To pass this prestudy a participant must have annotated 10 out of 15 examples correctly. Only around 10% of participants managed to pass the prestudy and the data was annotated by 14 annotators. Table 2 shows the number of instances created for each category of unseen data before and after human validation.[9] The results are consistent with our preliminary analysis with about 2/3 of the automatically extracted data being deemed correct by the annotators.

	E		E+P	
	B	A	B	A
# instances	4,167	2,126	3,800	1,312
# entities	7,801	4,038	11,264	4,078
# properties	57	53	394	296
# 1-triple G	3,725	1,917	374	176
# 2-triple G	334	172	326	127
# 3-triple G	53	27	647	295
# 4-triple G	9	1	755	256
# 5-triple G	3	0	782	240
# 6-triple G	43	9	916	218

Table 2: **KELM Extracted Subsets for English and Russian** Before (B) and After (A) human validation (E: graphs with unknown entities, E+P: graphs with unknown entities and properties).

Creating the Russian Benchmark. We create KELM-based benchmarks for Russian by automatically translating the texts of the English KELM benchmarks and manually verifying the resulting translations. For Machine Translation, we use the NLLB neural Machine Translation model (NLLB Team et al., 2022). For human validation, we hired four professional translators. As entities were shown to raise translation issues (Shimorina et al., 2019), we collected the Russian names of graph entities by querying DBpedia for their Russian label using the property 'rdfs:label' and provided the translators with (i) the English text from KELM, (ii) its translation into Russian and (iii) the Russian translation of the KG entities present in the input graph. Translators could copy and paste the NLLB translation and modify it afterwards. The translators also had the possibility to mention any kind of mistakes they notice.

Table 3 shows statistics on the changes introduced by the translators to convert the machine translated texts into valid Russian. To measure the differences between the two texts, we use the Levenshtein ratio. [10] We see a low similarity ratio indicating that, for Russian, machine translated texts needs correcting.

Translator	KELM Mean (STD)	WebNLG Mean (STD)
1	0.29 (0.15)	0.28 (0.16)
2	0.33 (0.15)	0.38 (0.16)
3	0.26 (0.15)	0.24 (0.18)
4	0.24 (0.15)	0.25 (0.16)
5		0.34 (0.16)
Total	0.28 (0.15)	0.30 (0.17)

Table 3: **Modification statistics between MT translations and final human translations for KELM and WebNLG test sets**. Levenshtein ratio distance mean and STD values for each translator separately and together.

Out of the 230 comments left by the translators, 214 concerned minor issues such as texts including + in front of positive numbers (the way they appear in the data). In two cases, the graph did not match a meaningful text and we removed either the whole instance or a triple from the graph. Finally, there were 14 instances where we modified both the English and the Russian sentence as these contained mistakes regarding the gender of a person (like a scientist was described as a man by default) or the lexicalisation of field specific terms (like 'taxon' in Biology).

5 Creating WebNLG Benchmarks for Russian

We derive two WebNLG Russian benchmarks from the WebNLG English test set by first selecting graphs with unseen categories or unseen entities

[9] One may notice the imbalance of the graph sizes for KELM-E. This is a consequence of a condition that all properties should be seen in WebNLG training/dev data. The more triples there are, the more properties there are in a graph and thus the less the possibility that all of them are seen.

[10] The Levenshtein distance indicates the minimum number of insertion, deletion or substitution of individual characters that are required to transform one sentence into another and the Levenshtein ratio normalises this distance by the length of the two sentences and inverts the score so that a perfect match will have a score of 1.0, and completely dissimilar strings will be assigned a value of 0.0 (LDistance: Levenshtein Distance, LRatio: Levenshtein Ratio): $LRatio(a,b) = 1 - \frac{LDistance(a,b)}{len(a)+len(b)}$

and second, translating the corresponding texts into Russian.

As explained above, English WebNLG differs from Russian WebNLG in that it covers 16 categories (vs. 9 for Russian) and the test set includes an Unseen Category and an Unseen Entities test set. To create an Unseen Category test set for Russian (WebNLG-C, 1,251 instances), we simply select from the English test set all instances which belong to the 7 categories not included in Russian WebNLG training and dev data. The second test set (WebNLG-E, 192 instances) consists of the instances that are from seen categories in Russian WebNLG train or dev set, but the entities are unseen.

These two subsets were then translated from English to Russian by 5 professional translators, who have Russian as a native language. As for the validation of the KELM translations, the translators were provided with the English text, the NLLB translation and the DBpedia Russian labels of the graph entities and again we observe a high ratio of changes introduced by the translators (Table 3).

Comparing the English texts to the corresponding graphs, the translators spotted a few errors (165 instances were highlighted out of the whole test set). Those errors include references to female scientists or politicians by he/him, subject and object interchanged in the text comparing to the KG data. We created a new version V3.1 of the WebNLG test data which integrates these corrections in the English version of the data and will be uploaded to the WebNLG website once this paper is published.

Table 4 summarises the created benchmarks indicating the number of test instances for each language and for each type of unseen data.

6 Assessing Generalisation

We evaluate current pre-trained Encoder-Decoders on our benchmarks. Since the best approaches in the 2020 edition of the WebNLG shared task were based on T5 or mT5 (Yang et al., 2020; Castro Ferreira et al., 2020), we consider various versions of this model fine-tuned on the WebNLG English/Russian training data. We also include in our evaluation the Control-Prefixes (Clive et al., 2022) model, a state-of-the-art model for KG-to-Text generation as well as the models for Russian submitted to the WebNLG 2023 Challenge (Cripwell et al., 2023). We evaluate the models using automatic metrics and run a qualitative analysis to identify the most common errors occurring when assessing current models on out of domain data.

Benchmark	Nb. of Instances	
	Russian	English
KELM		
KELM-E+P		
50/60	146	146
60/70	211	211
70/80	328	328
80/90	265	265
90/100	361	361
Total	1311	1311
KELM-E	2126	2126
WebNLG		
WebNLG-C	1251	N/A
WebNLG-E	192	N/A

Table 4: **KELM and WebNLG Unseen Benchmarks.** Number of (graph,text) pairs in each test set (E: Entities, E+P: Unknown Entities and Properties, X/Y: the min and max ratio of unknown elements, C: Unknown Category)

7 Quantitative Analysis

7.1 Models

English. We evaluate four models on the English benchmarks: the $T5_{base}$ model fine-tuned on the WebNLG 2020 training data for English ($T5_{ft}$); the $mT5_{base}$ and $mT5_{large}$ models fine-tuned on the WebNLG 2020 training data for English and Russian ($mT5_{base,ft}$, $mT5_{large,ft}$); and CP, a state of the art model for KG-to-Text generation (Clive et al., 2022)[11] which uses tasks-specific soft prompts (Control Prefixes, CP). We train this model for 40 epochs on WebNLG 2020 English training data with all the parameters provided by the authors and using their code.[12] When running the finetuned model on new KELM test sets, we pass categories (which are used as part of the prefix) all equalled to 1.

Russian. We also evaluate $mT5_{base,ft}$ and $mT5_{large,ft}$ fine-tuned on WebNLG Russian training data on the Russian benchmarks. In addition, we evaluate the mT0 pre-trained model (mT5 fine-tuned on crosslingual tasks, (Muennighoff et al.,

[11] https://paperswithcode.com/sota/data-to-text-generation-on-webnlg?p=control-prefixes-for-text-generation

[12] https://github.com/jordiclive/ControlPrefixes

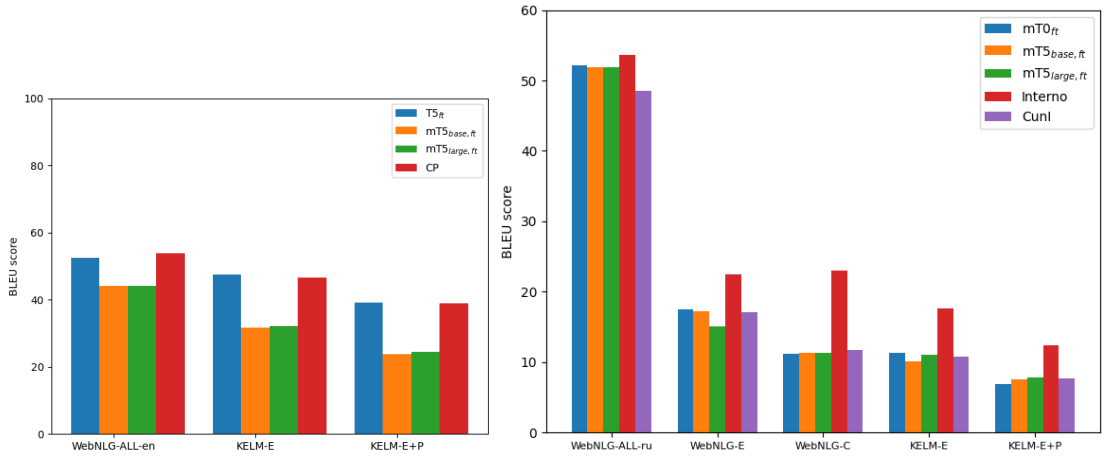

Figure 1: BLEU scores for each model on English (Left) and Russian (Right) Test Sets.

2023)) fine-tuned on the WebNLG training data for Russian (mT0$_{ft}$) and two models for Russian that participated in the WebNLG 2023 challenge. The first model is Interno, a model based on FRED-T5 (Full-scale Russian Enhanced Denoiser, 1.7M Parameters, (Zmitrovich et al., 2023)) and fine-tuned on WebNLG training data (Kazakov et al., 2023). We used the final checkpoints submitted to the WebNLG 2023 challenge. The second model is CunI, a mT5$_{base}$ model which was fine-tuned on multilingual data created by machine translating (using NLLB) WebNLG training data into Maltese, Irish Gaelic and Welsh and including the original Russian data (Kumar et al., 2023).[13]

7.2 Metrics

All models were evaluated using the WebNLG-toolkit[14] which includes the SacreBLEU implementation for BLEU (Papineni et al., 2002), the pyter implementation for TER6 (Snover et al., 2006), and the official implementations of chrF++7 (Popović, 2017) and BERTScore (Zhang et al., 2019).

7.3 Results

Figure 1 shows the BLEU scores for each model on each of the benchmarks. The results for the other metrics show similar trends, so they are not discussed in the paper but can be found in Appendix B.

Strong Degradation on the new Benchmarks. For all models and for both languages, we observe a strong degradation on our benchmarks with a drop in BLEU score with respect to the initial WebNLG test sets ranging from 5 to 20 BLEU points for English and 31 to 45 points for Russian. On English, the models that degrade least are the state-of-the art CP model and the monolingual T5 model fine-tuned on WebNLG. We observe a similar trend on Russian, where the degradation for the four multilingual models (mT0$_{ft}$, mT5$_{large,ft}$, mT5$_{base,ft}$, CunI) is worse than for Interno, a model based on FRED-T5 (Full-scale Russian Enhanced Denoiser), a monolingual model pre-trained on Russian. This suggests that multilingual models are more sensitive to out of domain data than monolingual ones.

Stronger Degradation on OOD Graphs. Comparing results on KELM and the WebNLG benchmarks (KELM-E/WebNLG-E and KELM-E+P/WebNLG-C), we find a stronger degradation on KELM benchmarks indicating that, even though there is a large overlap between DPedia and Wikidata properties and entities, models trained on DBpedia graphs and crowdsourced text do not generalise well to Wikidata graphs.

Stronger Degradation when both Properties and Entities are unseen. Unsurprisingly, we see that results are lower for graphs that contains both unseen properties and unseen entities (KELM-E+P, WebNLG-C) than only unseen entities (KELM-E, WebNLG-E).

Impact of the ratio of unseen elements. Figure 2 shows that performance mostly decreases as the ratio of unseen elements increases. There is a

[13] Unfortunately, we did not manage to reproduce the original results using the authors code(https://github.com/knalin55/CUNI_Wue-WebNLG23_Submission) and communicating with them. Possible difference: did not use the fp16 while it seems the authors used it (gpu available did not support it).

[14] https://github.com/WebNLG/webnlg_toolkit/

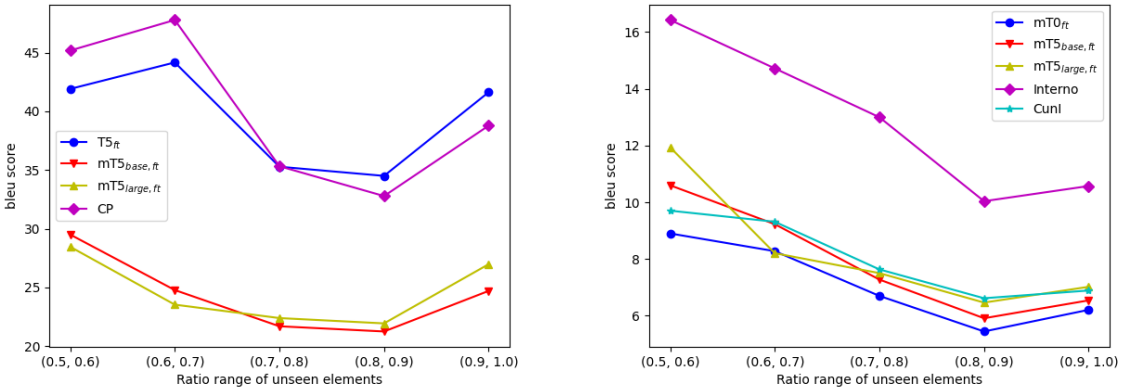

Figure 2: BLEU score for the different ratios of unseen elements (properties or entities) on English (Left) and Russian (Right)

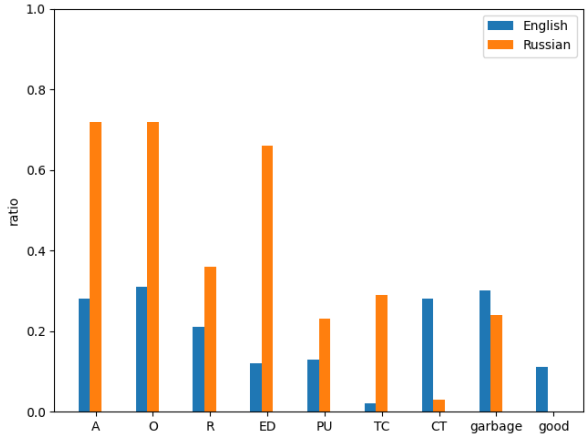

Figure 3: Error ratios per language

surprising peak at the 0.9/1.0 ratio, however. We conjecture that this is due to the high proportion of small graphs for this ratio (48% of these graphs are of size 1) which makes the generation task easier (cf. Table 7 in the Appendix).

We also see that, while for lower ratios of unseen elements, the mT5 base model (mT5$_{base,ft}$) outperforms the large one (mT5$_{large,ft}$), the inverse is true for ratios greater than 70%. This suggests that smaller models overfit the data. As the ratio of unseen elements is low, performance does not decrease too much as the remaining seen elements have been memorised by the model and can be generated correctly. Conversely, when the ratio is high, the advantage gained through memorisation of seen elements is reduced and performance decreases compared to larger models.

8 Qualitative Analysis

To get a better understanding of the type of errors made by generation models on OOD data, we run a qualitative analysis on the models outputs.

8.1 Error Annotation

For each model and each benchmark, we select the five instances with the lowest BLEU scores. This yields a total of 320 instances, 200 for Russian (8 benchmarks $\times$ 5 models $\times$ 5 instances) and 120 for English (6 benchmarks $\times$ 4 models $\times$ 5 instances). We then manually annotate the selected data for different types of errors including three error types previously used in the evaluation of KG-to-Text models (Belz et al., 2023) and six additional error types we found occurred in the data. Specifically, we identified the following 9 types of errors (The annotation was carried out by the first author who is a Russian native speaker).

Addition (A). The text contains information not present in the input graph.

Omission (O). The text misses information present in the input graph.

Repetition (R). The text has unnecessarily repeated parts.

Entity distortion (ED). An entity is mentioned in the generated text, but its name is partially incorrect. This can manifest in different ways for Russian and English. For Russian it includes entities copied over from the input data, entities mixing different scripts or just mistranslated. For English it mostly includes misspelling and incorrect numbers.

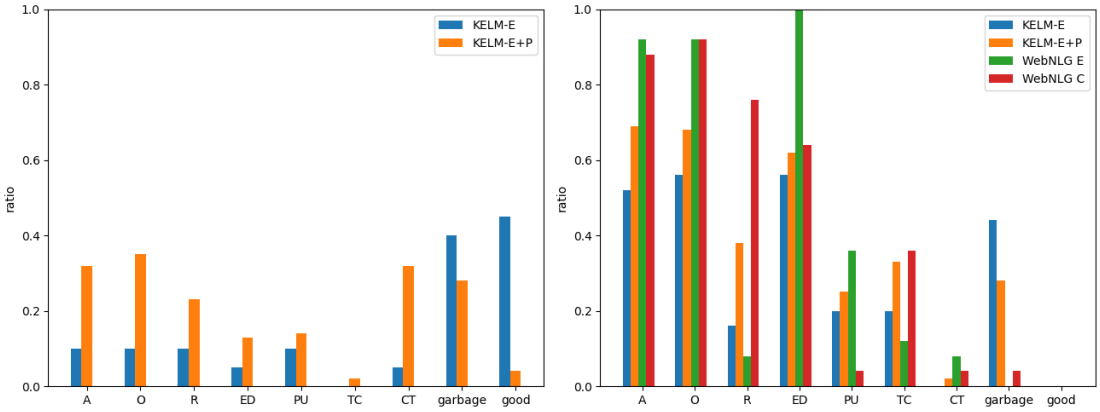

Figure 4: Error ratio per test set. English (Left) vs Russian (Right).

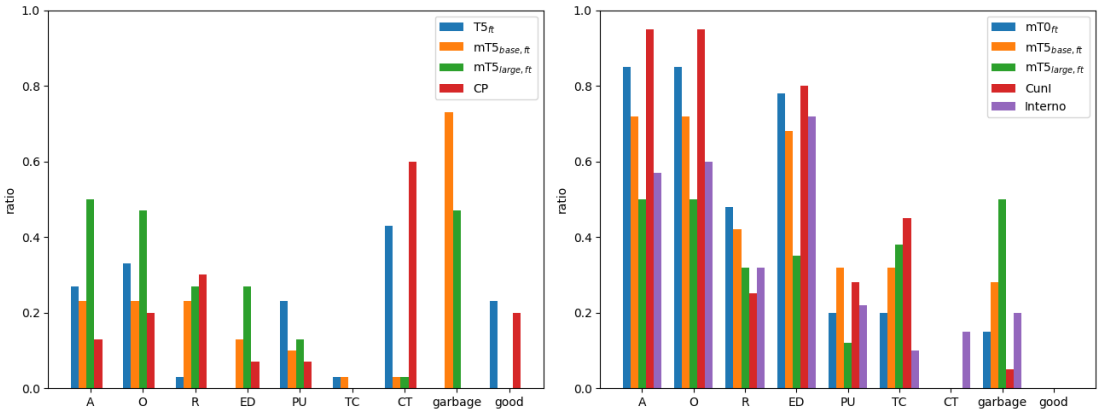

Figure 5: Error ratio per model. English (Left) and Russian (Right).

Property understanding (PU). The property is verbalised incorrectly (e.g., "instance of" is verbalised as *"is a part of"*).

Topic change. (TC) The text treats a property and its arguments as if they were from another topic for instance referring to buildings as if they were people and using expressions like *"was born on"* instead of *"was built in"*. This category differs from the "Property Understanding" category in that the lexicalisation of the property is correct out of context but incorrect for the given triple i.e., when taking its arguments into account.

Complex text (CT). The generated text is unnecessarily complex. This includes cases where each triple is verbalised but natural means of aggregation (ellipsis, coordination, pronouns) are not exploited resulting in unnatural text. E.g., *"Peter Slater (ornithologist) is a human and speaks, writes or signs in English. His given name is Peter."* rather than *"Peter Slater is an English speaking ornithologist."*). This error category also includes other over complications such as using *"is an instance of"* instead of directly saying *"is"*. This category is only assigned to cases which have neither additions nor omissions.

Garbage (G). Instances which consisted of just unrelated symbols or words which do not form any meaningful statements. If an instance is annotated as *Garbage*, no other annotation is assigned to it.

Good. Instances which in fact were good verbalisations of the input but received a low BLEU score because they paraphrased the reference text.

It is worth noting that one instance can contribute to several error annotations. E.g. *Property Understanding* often leads to one of the triples being not verbalised, and in this case we would also annotate the instance to have an *Omission*.

8.2 Error Analysis

Examples of each error types are given for both languages in the Appendix (Tables 8, 9 and 10). We also report error ratios per language, per model and per benchmark.

The error rate is markedly higher for Russian. Figure 3 shows a higher error ratio on Russian than on English overall highlighting a high level of degradation when the BLEU score is lowest. The high ratios for almost all error types indicate that the output texts contain multiple errors.

Domain change increases Topic Change errors. Interestingly, Figure 4 shows that topic change errors are more frequent on OOD data (KELM-E+P, WebNLG-C) highlighting the fact that neural models fails to adapt property verbalisation to the domain of discourse.

Custom Models show less errors overall. Figure 5 shows that for both Russian (Interno model) and English (CP model), custom models yield fewer errors overall than mT0 and mT5 fine-tuned on the WebNLG data.

9 Conclusion

We created challenging benchmarks for KG-to-Text generation into English and Russian, quantitatively demonstrated the effects of applying models trained on one distribution (e.g., WebNLG data) to a new distribution (e.g., unseen entities and/or properties) and identified nine error types which arise in this setting. The ability of existing generation models to generalise to OOD data is underexplored and we hope the benchmarks and evaluations we provide inspire further research on this topic, for instance under alternate KG-to-Text models.

Ethics Statement

During creation of the benchmarks we used Prolific to find annotators. Each annotator was provided with the annotation codebook. We did not gather any personal data during that process. We paid a rate of 10.5€ per hour. English-Russian translators were hired separately and paid according to their requested hourly rate. We use datasets (KELM, WebNLG) which are publicly available.

Supplementary Materials Availability Statement: We used the webnlg-toolkit[15] for evaluation and some of the model checkpoints available on that website. To avoid data contamination (Balloccu et al., 2024), the new test sets we developed will only be accessible through a web application which, given a file of generated output, will run all metrics available in the WebNLG toolkit and return the results to the user. This webapp is available at `https://webnlg-evaluation.loria.fr`.

[15] `https://github.com/WebNLG/webnlg_toolkit/`

Acknowledgements

We thank the anonymous reviewers for their feedback. We gratefully acknowledge the support of the French National Research Agency (Gardent; award ANR-20-CHIA-0003, XNLG "Multilingual, Multi-Source Text Generation"). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

References

Ankush Agarwal, Raj Gite, Shreya Laddha, Pushpak Bhattacharyya, Satyanarayan Kar, Asif Ekbal, Prabhjit Thind, Rajesh Zele, and Ravi Shankar. 2022. Knowledge graph - deep learning: A case study in question answering in aviation safety domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6260–6270, Marseille, France. European Language Resources Association.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

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.

Agnes Axelsson and Gabriel Skantze. 2023. Using large language models for zero-shot natural language generation from knowledge graphs. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 39–54, Prague, Czech Republic. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. 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.

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 363–382, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Liam Cripwell, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, and William Soto Martinez. 2023. The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 55–66, Prague, Czech Republic. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Maxim Kazakov, Julia Preobrazhenskaya, Ivan Bulychev, and Aleksandr Shain. 2023. WebNLG-interno: Utilizing FRED-t5 to address the RDF-to-text problem (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 67–72, Prague, Czech Republic. Association for Computational Linguistics.

Nalin Kumar, Saad Obaid Ul Islam, and Ondřej Dušek. 2023. Better translation+ split and generate for multilingual rdf-to-text (webnlg 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 73–79.

Teven Le Scao and Claire Gardent. 2023. Joint representations of text and knowledge graphs for retrieval and evaluation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 110–122, Nusa Dua, Bali. 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.

Simon Mille, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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. *Journal of machine learning research*, 21(140):1–67.

Anastasia Shimorina, Elena Khasanova, and Claire Gardent. 2019. Creating a corpus for Russian data-to-text generation using neural machine translation and post-editing. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 44–49, Florence, Italy. 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 *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian.