Inroads to a Structured Data ↔ Natural Language Bijection and the role of LLM annotation

Blake Vente
rv2459@columbia.edu

Abstract

I find evidence affirming the theory that using multiple training objectives jointly with sequence-to-sequence transformer language models can improve performance on some subtasks. In particular, when training between data and natural language sentence representations, I observed that the multi-task generalist model outperforms the specialist model with a F_1 of 0.771 up from 0.692, likely due to some form of cross-task knowledge generalization. This suggests that the same network "re-using" the same data in a different way may lead to higher performance in some subtasks. However, the inverse task alone is likely not enough to outperform all subtasks in all situations. And, I find furthermore that a multi-task t5-small fine-tuned on 33 percent LLM annotations does not perform substantially better or worse than the same model without the synthetic data.

1 Introduction

1.0.1 Motivation

It would be an understatement to say there has been an explosion in interest in Large Language Models (LLM's) for assisting with knowledge work. At the same time, we are grappling with the factual confabulation problem (Gabriel et al., 2021; Kryscinski et al., 2020).[1]. To address this, have proposed Retrieval-Augmented generation (RAG), placing unstructured documents into the context windows of Large Language Models. Still, there is a relative dearth of researching structured queries for RAG from structured sources compared to unstructured sources (Li et al., 2022; Shuster et al., 2021). It may one day be possible for smaller language models to match or best larger models for factual data recall tasks by iteratively querying databases of facts with their sources. This work is a first step towards Pre-trained Language Models which add structure to documents. Further aspirational use-cases follow in Appendix F.[2]

2 Related Work

Multitask Training Given sufficient learning capacity, it is possible for a language model to score better on all tasks by increasing the number of tasks via a mechanism called "co-training transfer." (Aribandi et al., 2021). The authors present ExT5 with the gargantuan figure of 107 training tasks and evaluate how and when a multi-tasking model can outperform a model of the same size with no additional data from that particular dataset. This work inspired the choice of WikiBio as an appropriate related task.

Freeze No Layers The current state-of-the-art methodology "control prefixes" reaches a BLEU score of .67 on seen entities and a .61 overall Clive et al.. "Control prefixes" are control signals that are directly appended to the hidden states of the network to guide generation while using a frozen pre-trained language model (PLM). This work recommends against freezing any layers for the propagation of prompt prefixes in control Raffel et al. (2022).

Semantic Parsing For Semantic Parsing (sentence-to-data and s2d used as direct synonyms in this work), the current state of the art on the WebNLG+ dataset Dognin et al. (2021) with F_1 score of .723. This work very innovatively frames text generation as a multi-step decision process and discusses adaptations to use non-differentiable evaluation metrics as a reinforcement penalty to guide generation. For this task, F_1 score is derived from framing semantic parsing as a closed-world classification task where an unordered set of RDF triples is generated. The WebNLG corpus was

[1] Used as a synonym in lieu of "hallucination" to avoid anthropomorphizing, per (Edwards, 2023)

[2] The repository can be found at https://github.com/rvente/nlgs-research/, which contains links to published model weights.

originally released as part of a competition, and the work Castro Ferreira et al. (2020) compiles the results and summaries of many approaches. Notably, there is support for multitask learning as an approach with the bt5 network. In particular, Agarwal et al. (2020) used cross-lingual multitasking for English and Russian, with a final Resulting F_1 score of .877.

Synthetic Data Generation Acording to Axelsson and Skantze many language models can perform data-to-sentence generation by default, but despite high fluency, ChatGPT earns a paltry 0.424 BLEU, which speaks to the limitations of automatic evaluation. Work by Shin and Durme (2022) shows that using longer beam lengths increases accuracy on two d2s corpora, Overnight and SMCalFlow. Separately, work by (Tang et al., 2023) shows promising results from generating synthetic annotations for the BioCreative VCDR corpus of clinical text (F_1 from 0.2337 to .6399 for named entity recognition; and F_1 from 0.7586 to 0.8359 percent for relation extraction). Finally Hsieh et al. (2023) uses PaLM-generated rationales as a proxy objective for smaller pretrained language models whose primary task is natural language inference on the Stanford Natural Language Inference (SNLI) and Adversarial Natural Language Inference (ANLI) corpora.

3 Data

The primary corpus for this work is the "bi-directional WebNLG+" (also called WebNLG+ 2.0 or WebNLG2020) variant of the widely-used WebNLG corpus introduced in Gardent et al. (2017). The "bi-directional" descriptor alludes to the two sub-tasks: RDF-to-sentence and sentence-to-RDF. Castro Ferreira et al. (2019) calls these tasks generation (data-to-sentence, d2s) and semantic parsing (sentence-to-data, s2d) respectively. RDF triples contain three terms, with entities on either side surrounding a relation. The relations themselves have an extremely imbalanced distribution, with most relations occurring extremely rarely and few relations occurring extremely frequently. This can be seen in Figure 1d.

Examples of individual training examples can be found in Figure 5. The average length in tokens can be found in 1a. Each record of the corpus has exactly one set of RDF triples and on average 2.66 of natural language translations of that particular triple set. In turn, RDF triple set has an average of 2.9 individual RDF triple items with a standard deviation of 1.5. This corpus has 12,876 total records in the training set, 1619 in validation, and 1600 in the test set. The sentences of Wikibio are much longer on average: with 526.48 characters per sentence.

The d2s task inputs RDF triples (structured **d**ata) as and outputs natural language text (**s**entences). Conversely, the s2d task inputs natural language sentences and outputs structured data. I use these as direct synonyms of generation and semantic parsing respectively. I performed basic data cleaning on the corpus, including Unicode to ASCII remapping.[3]

4 Methods

This work establishes baseline models, referred to as the "specialists" that are fine-tuned on a single task and compares it to the multi-taskers, "generalists". To this end, I fine-tune pre-trained versions of the t5 series from Su et al. (2021) three sub-tasks: data-to-sentence, sentence-to-data, and multi-tasking. I fine-tune from the plain model found at HuggingFace[4]. For the multitasking model, an arbitrary control prompt specifying the task is prepended to the input, `d2t 0:` and `t2d 1:`. The single-task variants do not need this additional information to specify the task. The training tasks are interlaced, namely, the tasks alternate ABAB until all training examples are exhausted. For all tasks, structured data must be serialized for insertion into the context window. For data-to-sentence, the RDF triple terms are joined with vertical bars `|`, and each triple ends with a `;`. For sentence-to-data, the natural language sentence is taken as input and decoded into the serialized RDF representation.

This work also incorporates Google's PaLM Large Language Model (Chowdhery et al., 2022) for automatic data annotation. I use the text-bison version of Google PaLM 2 via Google Cloud. PaLM is an instruction-tuned large language model, so I will rely on in-context learning and prompt control rather than explicit training. Due to time constraints, I leave it to future work to rigorously evaluate PaLM on the sentence-to-data task in WebNLG. This work only uses the LLM as an annotator, outputting WebNLG-style triples for each WikiBio record.

[3]The practice of replacing underscores with spaces to obtain a more token-efficient representation was substantiated in my midterm, omitted here due to space constraints.

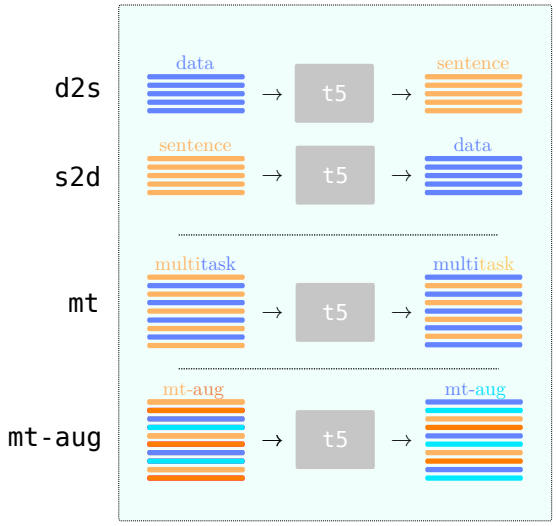
[4]`https://huggingface.co/t5-small`

	WebNLG		WikiBio
	Sentence	RDF	Sentence
mean	310.68	139.54	526.48
std	171.69	83.79	593.25
min	22.00	22.00	6.00
25%	174.00	71.00	183.00
50%	291.00	130.00	318.00
75%	421.00	190.00	627.00
max	1191.00	657.00	56698.00

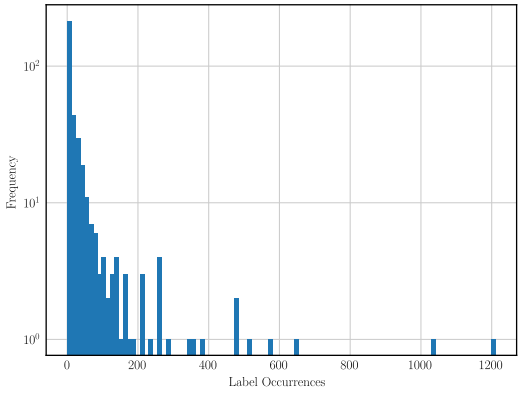
(a) Length in characters for the various corpora. These statistics are reported before any preprocessing is done to the text.

Corpus	WebNLG	WikiBio
train set	12876	582659
validation set	1619	72831
test set	1600	72831
Σ	16095	728321

(b) Number of records for each corpus by partition set.



(c) **Model Training Scheme** There are three treatments are: (1) specialists: data-to-sentence and sentence-to-data (2) multi-task and (3) synthetic data-augmented variant of 2.



(d) Some relations such as `country` and `location` have more than a thousand occurrences. But there is a long tail of relations that only occur once, including `nearestCity` and `currentTeam`.

Figure 1: Corpus details and Training Scheme

5 Experiments

The Huggingface wrapper to the reproducible sacreBLEU library from (Post, 2018) computes BLEU-4 scores as specified by in (Papineni et al., 2002). Comparison was case-insensitive. Huggingface's `evaluate` library computes BERTScore (Zhang et al., 2020) with `distilbert-base-uncased`, a distilled variant of the English Bert model. Likewise, `evaluate` computes the RougeL score, the Longest Common Subsequence Rouge Score (denoted RougeL), wrapping the Google Open Source implementation of Rouge.

The experiments adhered to the standard train, development, and test splits of the corpus. As the baseline model, I fine-tuned the Huggingface `t5-small` and `t5-base` networks with an effective batch size of 64 and 32 respectively. Both networks were trained with a learning rate of $2 \cdot 10^{-4}$ for 5 epochs. The shortest training time for any network was about 30 minutes and the longest took about 1.5 hours (no counting evaluation). There was a wide variance of decoding time during the predictions in the validation set (more on this in Appendix 3). Decoding was performed with 4 beams, temperature was reduced from the default of 1.0 to 0.9, and top-k filtering was kept at the default of 50 tokens. The first treatment is the multitask training objective, where the model must predict perform s2d and d2s alternating in the records, and the second treatment is the same as the first, but also incorporates LLM annotated WikiBio data randomly inserted into the training set, as denoted in

1c.

Training proceeded on a 24 GB Nvidia RTX 3090. At the time of training, t5-base with a batch size of 32, `nvidia-smi` reports 19289 mebibytes (MiB; ≈ 20.2 gigabytes) of Video RAM (VRAM) in use. This figure slowly grows as memory fragmentation occurs during the training process. For example, by epoch 1.4 it grew by 1500 MiB which is about 8 percent of VRAM consumption at the start of training. The amount of memory leaking is linear. Decreasing batch size to 16 resulted in a divergence of training loss within 5 epochs on t5-base. To resolve this, I used `gradient_accumulation_steps` for a larger effective batch size even with a smaller in-memory batch size [5].

6 Results

In this case, taking t5-small and training it with the multi-task objective caused it to perform higher on the d2s metrics F_1 (from 0.692 to 0.771) and Edit Distance (from 16.854 characters to 16.16), compared to the same model trained on just a single objective. At the same time, it didn't substantially change BERTScore, and resulted in a slight decrease in Rouge (from 0.745 to 0.733), and BLEU (from 0.641 to 0.618). This might show signs of *multi-task generalization*, but only in one direction. One conjecture into the underlying mechanism could be that the model learns vocabulary that is re-used when computing labels. This would explain why the converse isn't true: RDF label vocabulary is a subset of the text vocabulary.

However, in these same circumstances, the multi-task t5-base model was lower than the respective specialist models in every metric. Its closest was in BERTScore where performance only fell slightly (down to 0.945 from 0.957). Taken on its face, this might suggest there exists an ideal ratio between training set size, fine-tuning size, and model size similar in spirit to Palm's compute optimal scaling hypothesis (Hoffmann et al., 2022). However, much more research is needed to add confidence to this finding, explored more in the Limitations portion.

As an aside, I observed that pre-pending the task may not be strictly necessary because the model was able to identify the task automatically, probably from the vertical bars that delimit the RDF triples.

[5] https://huggingface.co/docs/accelerate/usage_guides/gradient_accumulation

LLM Annotation I requested 5,000 annotations from PaLM using the text-bison API. About 400 were content-filtered: by default, the API filters out model outputs about protected groups. About 100 contained malformed expressions. [6] The 4543 remaining records were processed in the same manner as the WebNLG corpus. Qualitatively, the model appeared adequate to perform semantic parsing, and in the samples I have observed, did not add any information not originally present in the prompt. However, t5-small-aug did not perform better than t5-small. In fact, the augmented variant performed slightly worse in all tasks. Perhaps this was due to not enough learning capacity in the t5-small base model. Perhaps idioms of WikiBio were different enough as to slightly harm performance. More work is needed for a full explanation.

7 Error Analysis

"Repetition loops" as in (Xu et al., 2022) refers to when the model repeats the same few sentences over and over again. I observed this "hallucination" occur, even in the largest set of models trained `t5-base-mt`. Curiously, it was isolated to a particular word Palatul, Romainan for "palace" 3. This behavior diminished but was present in some of the trained models nonetheless. One trained model had a "Palatul" glitch token. In Figure 3, each row is an example of t5-base-mt falling into repetitive generation cycles even when the `no_repeat_ngrams` parameter is set to 3. Ultimately it is the parameter of `max_length` that terminated generation. To reduce validation length, one can set this value low.

The errors show the existence of some "false penalties" where the model's outputs are sensible but not recognized by automatic evaluation. As seen in Figure 7, BLEU does not take into account sentence variations or semantic similarity. In those records (from d2s-t5-small) the BLEU was than 0.15 In record 222, we can see a penalty for using 30.0 g instead of 30 grams. We also see penalties from differential placement of relative clauses (in this case "whose", "who is from spain"). Across these samples, the BERTScore is 0.893. This casts further doubt into these forms of automatic evaluation.

In Figure 8, I report plots on the worst performers on semantic parsing. In all variants after nor-

[6] To be space-efficient, my prompt is available in my code repository under the `palm` folder. The pickle file for the annotated records is `wikibio_llm_annot.pkl`

	BLEU↑	BERTScore ↑	RougeL↑	F_1↑	Edit Distance ↓
	data-to-sentence			sentence-to-data	
t5-small	0.641	0.953	0.745	0.692	16.854
t5-base	0.671	0.957	0.767	0.928	15.489
	multi-task				
t5-small	0.618	0.949	0.733	0.771	16.16
t5-base	0.602	0.945	0.718	0.887	15.711
	multi-task llm-augmented				
t5-small	0.610	0.948	0.730	0.754	16.359

Figure 2: This table reports the results of the six experiments run on the WebNLG corpus. The identical standard test set was used for all evaluation $N = 1600$. The figures reported are not directly comparable to prior work due to cleaning and preprocessing.

malizing by the prevalence in the training set, I can confirm that `SportsTeam` was disproportionately difficult for the model, even though these samples were shorter than average. Due to time constraints, it can be left to future work to investigate further, especially the question of t5-base-aug.

8 Conclusions, Limitations, and Future Work

The key finding of this work is that even with no additional data or an increasing model size, learning the inverse task may increase performance for some sequence-to-sequence sub-tasks. This finding is significant because it upholds the value of multi-tasking, even in a two-task context with no additional data. But this is not reliable as only some metrics increased while others decreased. Scaling up the model was a more reliable way to gain more performance.

Limitations In this work, every model was trained with the same control value of 5 epochs through the training set, so every model trained on the same number of records the same number of times. However, future work should certainly take into consideration the effect of training each and every network fully to convergence, and experimenting with different training parameters. In this work, I also didn't perform statistical significance tests. Given more time, I would train each network multiple times, or use several independently seeded test set generations.

In this work, long entity names were negatively correlated to performance on both sub-tasks. Since the task is framed as a sequence-to-sequence problem, the individual decoding errors add up. It's possible to "compress" the text form of the entity so that long entity names don't "distract" from the true purpose of the model. In particular, very long entity names could be bound to separate short-names and then unbound later. To illustrate one intuitive compression scheme Appendix B shows an example. One can see that this saves on token count compared to repeating the entity name without losing any information. However, it may require the network to learn to bind a name to a variable and use it appropriately. In this work, this optimization was not explored, but future work may consider this.

The "Palatul" issue shows that the "Hallucination problem" still requires addressing. I conjecture that one approach to mitigate this is defining a task to the index of the spans themselves for entity names, (just as an image segmentation model might predict bounding boxes around objects in an image). By forcing a model to work with indices directly, extracting a sequence not present in the input would be structurally impossible. Prior work shows that this method is feasible in principle (Subramanian et al., 2021), but much further research is required. To be specific, instead of the probabilistic mechanism of decoding input tokens from encoded representations through the network, the task would entail directly predicting the index of the substrings that comprise the terms of each RDF triple.

There is still no defacto standard normal representation for all words, so "Footballer" and "Soccer player" may have different RDF labels, but this

isn't desirable in general for automatic evaluation as seen in E.1. Even an informal notion of *bijection* between Natural Language and Structured data demands solutions to the above problems. It would be a daunting task to approach this, but families of models such as this may provide a path forward.

A F_1 measure definition

For rigor's sake, F_1 score is not used in its usual sense, so it is fruitful to formally extend the definition of F_1 for sets of open-vocabulary labels. There is no defacto implementation or pseudocode between all works reporting F_1 score for semantic parsing (s2d) task. Let P be the set containing predictions in the form of RDF sequences from the model. Let G be the set containing the ground-truth references in the same format. Let $\mathbf{hm}(a, b)$ be the harmonic mean of integers a and b, and let $\mathbf{hm}(0, x) = \mathbf{hm}(x, 0) = 0$ for all x. And let $|X|$ denote the cardinality of set X.

$$
\begin{aligned}
&\textit{integer } \text{TP} \leftarrow |P \cap G|;\\
&\textit{integer } \text{FP} \leftarrow |P - G|;\\
&\textit{integer } \text{FN} \leftarrow |G - P|;\\
&\textit{decimal } \text{PREC} \leftarrow \text{TP}/(\text{TP} + \text{FP} + \varepsilon);\\
&\textit{decimal } \text{RECL} \leftarrow \text{TP}/(\text{TP} + \text{FN} + \varepsilon);\\
&\textit{decimal } F_1 \leftarrow \mathbf{hm}(\text{prec}, \text{recl});
\end{aligned}
$$

In the following code, P is the first argument and G is the second. The strings in the sets are always computed by case-insensitive, whitespace-insensitive match.

```
f_measure(set("a"), set('a')) == 1
f_measure(set("ab"), set('a')) == 2/3
f_measure(set() , set('a')) == 0
```

B Semantics Preserving Compression Schemes

```
let [A] = "The Spirit of Christmas Yet To Come";
[A] | appears in | A Christmas Carol
[A] | is a | fictional character
[A] | is a | ghost
[A] | created by | Charles Dickens
[A] | appears before | Ebenezer Scrooge
```

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.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Prakash Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *CoRR*, abs/2111.10952.

Agnes Axelsson and Gabriel Skantze. 2023. Using large language models for zero-shot natural language generation from knowledge graphs.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2019. The 2019 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 2rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 54–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation.

Pierre L. Dognin, Inkit Padhi, Igor Melnyk, and Payel Das. 2021. Regen: Reinforcement learning for text and knowledge base generation using pretrained language models.

C "Hallucination" Example: The "Palatul" Glitch Token

predictions	palatul_count
Austin is part of Hays County,Texas.	2028
Arlington is part of Tarrant County,Texas.	2028
The Acharya Institute of Technology	2034
The address of Amdavad ni Gufa is "	2025
alatul Ryan is the leader of the United States, where English is spoken.	2025
The ground of AFC Blackpool is located in Blackpool which is led by Gordon Marsden.	2020
Austin is part of Hays County,Texas,which has San Marcos as its county seat.	2018
Augustus Pugin was born in Bloomsbury, Ireland. He was the architect of Adare Manor, of Westminster,	2013
The 1 Decembrie 1918 University is located in Bucuresti Bucu	2008

Figure 3: A sample of the "Palatul" curse, generated from the `t5-base-mt` variant during training. Each string starts with a typical generation and diverges into this repetitive generation cycle. This behavior occurred with generation parameter `no_repeat_ngram=3`. I observed that values larger than this caused the model to backtrack for hours at certain points during training. This points Future work might explore this problem in detail. This happened 9 times in the entire 1600 record test set during training of one model, and subsequently, the problem diminished without a clear cause. Similar behavior has been documented as "Glitch Tokens" https://www.lesswrong.com/tag/glitch-tokens

B Edwards. 2023. Why chatgpt and bing chat are so good at making things up. *Ars Technica. https://arstechnica.com/informationtechnology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them*.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. 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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

D Data Samples

D.1 Arbitrary WebNLG Records

```
241 In Mexico, the spoken language is Spanish.
    Spanish is the language spoken in Mexico.
    The language of Mexico is Spanish.
242 One of the languages used in the Philippines is Arabic.
    Arabic is a language spoken in the Philippines.
    One of the languages in Philippines is Arabic.
    Arabic is one of the languages spoken in the Philippines.
243 Shumai is a variation of the dish Siomay.
    Siomay and Shumai are variations of the same dish.
244 Native Americans in the United States are one of the ethnic groups of the
    country.

241 Mexico | language | Spanish_language
242 Philippines | language | Arabic
243 Siomay | dishVariation | Shumai
244 United_States | ethnicGroup | Native_Americans_in_the_United_States
```

Figure 4: Some arbitrarily chosen samples of the sentence form of the data (top) with the associated raw data in triple form (bottom). Each entry that appears on a new line is one of the valid options for the data-to-text task. This means that BLEU score will acknowledge each of the variants. The natural language text is above and each numbered line is paired with the structured RDF triples below.

D.2 Arbitrary WikiBio Text Records

```
john chubb -lrb- 1816 -- 1872 -rrb- , was an english locksmith and inventor . he
    wrote an important paper on locks and keys , and was awarded the telford medal .
mary kendall browne -lrb- june 3 , 1891 -- august 19 , 1971 -rrb- was the first
    american female professional tennis player , a world no. 1 amateur tennis player
    , and an amateur golfer . she was born in ventura county , california , united
    states .
nicholas phillip ebanks -lrb- born 27 june 1990 -rrb- is a caymanian footballer who
    plays as a defender . he has represented the cayman islands during the 2010
    caribbean championship and world cup qualifying matches in 2011 .
warren archard luhning -lrb- born july 3 , 1975 -rrb- is a retired canadian
    professional ice hockey winger .
```

Figure 5: Further arbitrarily chosen samples from the text column of WikiBio. These are pre-tokenized and have ASCII encoding for individual characters. The whole sequence is lowercased and `-lrb-` denotes "left round bracket".

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivek Subramanian, Matthew Engelhard, Sam Berchuck, Liqun Chen, Ricardo Henao, and Lawrence Carin. 2021. SpanPredict: Extraction of predictive document spans with neural attention. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

E Model Outputs

E.1 Qualitative Observations on False penalties

1. Transitive relations whose arguments are swapped are not captured
2. Synonymy/ reasonable alternatives not accounted for
3. Equivalent formulations not accounted for
4. Units and formulation
5. Inconsistent schema/synonymy

```
1 P: Christian Burns|associated band/associated musical artist|Andrew Rayel
  A: Andrew Rayel|associated band/associated musical artist|Christian Burns

2 P: California|stone|Benitoite
  A: California|gemstone|Benitoite

3 P: Al Kharaitiyat SC|league|Qatar Stars
  A: Al Kharaitiyat SC|position|Qatar Stars League

4 P: Andrews County Airport|runway length|896
  A: Andrews County Airport|runway length|896.0

5 P: Atlanta|leader name|Kasim Reed
  A: Atlanta|leader|Kasim Reed
  P: United States|leader name|Barack Obama
  A: United States|leader|Barack Obama
```

Figure 6: (P = predicted, A = actual; examples extracted from t5-small trained to 5 epochs)

E.2 Low BLEU scores

```
198 | ["April O'Neil was created by Kevin Eastman.", "Kevin Eastman is the creator of April O'Neil."]
    > Kevin Eastman created April ONeil.
222 | ['Barny cakes can be served in 30 gram sizes.', 'Serving size for the Barny cakes is 30.0g.', 'The serving size of Barny
      cakes is 30.0g.']
    > Barny cakes have a size of 30.0 g.
229 | ['Bionico can be varied by using cottage cheese.']
    > Cottage cheese is a variation of Bionico.
288 | ['Abdulsalami Abubakar ended his career on 1999-05-29.']
    > Abdulsalami Abubakar's career ended on 29th May 1999.
310 | ['Allan Shivers started his career from January 21, 1947.']
    > Allan Shivers began his career on 21 January 1947. |
419 | ['Alan Frew is a rock musician, which includes fusion and Bhangra styles.', "Alan Frew's genre is Rock music of which
      bhangra is a fusion of rock.", "Alan Frews' musical genre is rock music and a type of rock music fusion is Bhangra."]
    > Alan Frew performs rock music which has a fusion genre called Bhangra.
438 | ['Al Anderson of NRBQ is a country musician in which genre the banjo features.', 'Al Anderson of NRBQ performs country
      music which is a genre of music which uses the banjo.', 'Al Anderson (NRBQ band) performs country music, in which the
      banjo is one of the instruments.']
    > Al Anderson, a member of the NRBQ band, performs country music. Banjo is a musical instrument of country music.
551 | ['Baked Alaska comes from the country of France and one of the ingredients is sponge cake.', 'Baked Alaska (France) uses
      sponge cake as an ingredient.', "France's Baked Alaska includes the ingredient, sponge cake."]
    > Sponge cake is an ingredient in Baked Alaska which is from France.
570 | ['Bionico requires granola as one of its ingredients and can be found in Guadalajara.', 'Granola is a required
      ingredient of the Guadalajara regional dish, Bionico.', 'Bionico, which contains granola, can be found in Guadalajara.']
    > Granola is an ingredient in Bionico which comes from the Guadalajara region.
755 | ['Alan Martin, whose club is Motherwell FC, played for Accrington Stanley FC who have their ground in Accrington.', "
      Alan Martin's football club is Motherwell FC and he has also played for the Accrington based club Accrington Stanley."]
    > Alan Martin is a footballer for the Accrington Stanley F.C. club which is located in Accrington.
809 | ['The epoch date of 1097 Vicia, which had 1928 PC as its former date, is 2006.12.31. Vicia has a periapsis measurement
      of 279142000000.']
    > 1097 Vicia, formerly known as 1928 PC, has an epoch date of December 31st 2006. It has a periapsis of 279142000000.0. |
885 | ['Found in Mexico, the food, Bionico (with granola as an ingredient), is served at the dessert course.', 'Bionico is
      served as a dessert course. It is found in Mexico and requires granola as an ingredient.']
    > Bionico is a dessert from Mexico and contains granola.
```

Figure 7: A sample of BLEU score < 0.15 generations from d2s-t5-small. The Training set recordId starts each listing, followed by a vertical bar, then the list of ground truth translations, and finally > begins the generation of the model.

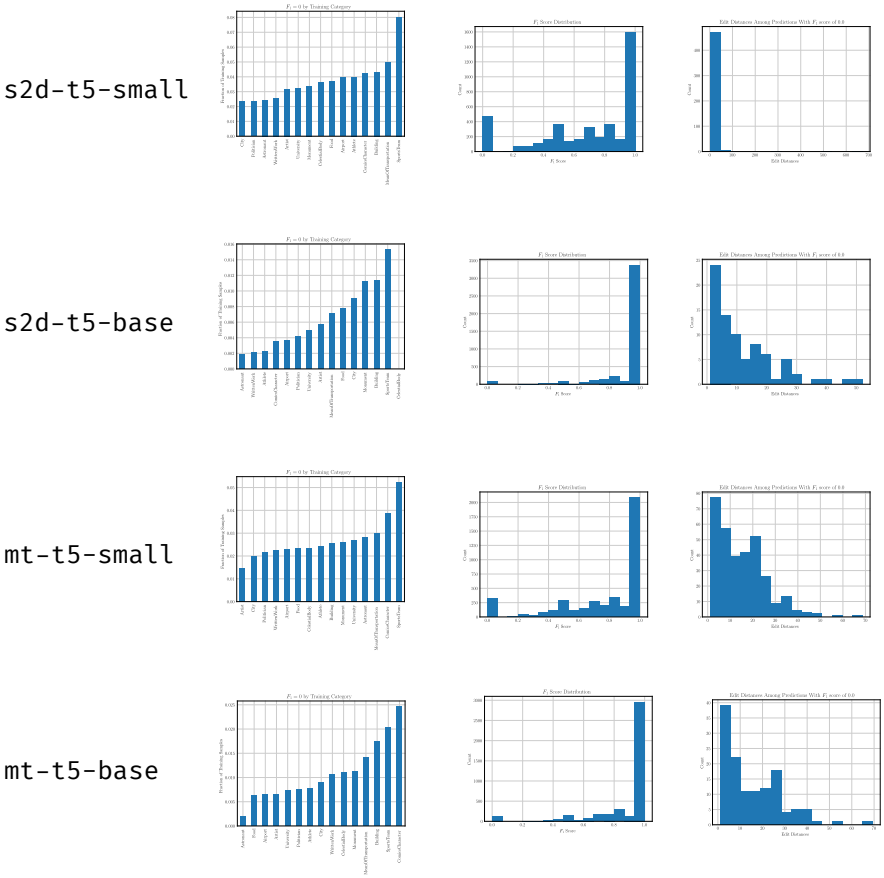


Figure 8: Performance on the sentence-to-data task by network size and training scheme. The first column represents the records that were secured an F_1 of 0 by their training category. The second shows the overall F_1 score distributions. The final column shows the edit distance between the terms and their expected values. For the s2d-t5-small variant, the Edit Distance plot does not show it well, but the horizontal axis still included a single a single record that resulted from repetitive generation cycles.

F Further Motivating Use-cases for Text-to-Data

To plan further developments to this task, future work may consider the following use-cases. All of these use-cases further substantiate the need for standard, normalized data formats, which I consider the chief limitation faced by this task moving forward.

1. Pipeline generation for Summarization: it's possible to break up a text into a set of facts and then rank the facts in terms of importance, de-duplicate them, sort them, and convert them back into a summary. Viewed this way, this holds promise for long-form summarization tasks.

2. Summarization factuality evaluation: if there exists in the future, some exhaustive normal form capturing entities, relations between them, and their evolution over time, one way to evaluate factual fidelity would be to compute this "factual decomposition." of the source text, its summary, and then compute the overlap. A faithful summary would be a proper subset of the facts in a source text.

3. Editing Wikipedia may be more beginner-friendly than editing on Wikidata. A system that extracts facts from new Wikipedia contributions and auto-syncs them with Wikidata (and the inverse) would be fruitful for the mission to keep the two platforms in knowledge parity.

4. Corroborating evidence and finding factual inconsistencies between various sources, which may be extended into automated fact-checking.

Technologies, pages 5234–5258, Online. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining?

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.