WikiGraphs: A Wikipedia Text - Knowledge Graph Paired Dataset

Luyu Wang† and Yujia Li† and Ozlem Aslan and Oriol Vinyals
†Equal contribution
DeepMind, London, UK
{luyuwang, yujiali, ozlema, vinyals}@google.com

Abstract

We present a new dataset of Wikipedia articles each paired with a knowledge graph, to facilitate the research in conditional text generation, graph generation and graph representation learning. Existing graph-text paired datasets typically contain small graphs and short text (1 or few sentences), thus limiting the capabilities of the models that can be learned on the data. Our new dataset WikiGraphs is collected by pairing each Wikipedia article from the established WikiText-103 benchmark (Merity et al., 2016) with a subgraph from the Freebase knowledge graph (Bollacker et al., 2008). This makes it easy to benchmark against other state-of-the-art text generative models that are capable of generating long paragraphs of coherent text. Both the graphs and the text data are of significantly larger scale compared to prior graph-text paired datasets. We present baseline graph neural network and transformer model results on our dataset for 3 tasks: graph → text generation, graph → text retrieval and text → graph retrieval. We show that better conditioning on the graph provides gains in generation and retrieval quality but there is still large room for improvement. ¹

1 Introduction

Parallel datasets that pair data from different sources and modalities have enabled large amounts of research on cross modality learning. Paired image-caption datasets enable models to describe visual scenes in natural language (Lin et al., 2014; Vinyals et al., 2016), paired streams of speech and transcription data makes it possible to train speech recognition systems (Garofolo et al., 1993; Panayotov et al., 2015) or text-to-speech synthesis models (Oord et al., 2016), and parallel corpus of text in different languages enable learned machine translation models (Barrault et al., 2020).

¹The data and the code to reproduce our baseline results are available at https://github.com/deepmind/deepmind-research/tree/master/wikigraphs

Figure 1: Illustration of a pair of Wikipedia article and the corresponding knowledge graph in our dataset.

We present a new dataset of Wikipedia text articles each paired with a relevant knowledge graph (KG), which enables building models that can generate long text conditioned on a graph structured overview of relevant topics, and also models that extract or generate graphs from a text description.

There has been many prior efforts trying to build datasets for learning graph → text generation models (Jin et al., 2020; Gardent et al., 2017; Lebret et al., 2016). However, existing graph-text paired datasets are mostly small scale, where the graphs tend to have 10-20 or even less nodes, and the text typically only contains one or a few sentences. This represents a significant contrast with the state-of-the-art text generation models (Dai et al., 2019; Brown et al., 2020), which can already generate very fluent and long text that spans thousands of tokens over multiple paragraphs.

We attempt to bridge this gap, with the goal of advancing the state-of-the-art graph → text generation models, graph representation learning models and also text-conditioned graph generative models. Each text document in our dataset is a full-length Wikipedia article, and we pair each of them with a KG that are significantly bigger than prior datasets of similar nature and includes much richer information. Hand labelling text articles with KGs is expensive and not scalable (Jin et al., 2020), there-
fore we utilize an existing and established knowledge base, Freebase (Bollacker et al., 2008), and designed an automated process to extract a relevant subgraph from it for each Wikipedia article. To make the text generation results on our dataset directly comparable to the state-of-the-art, we chose the set of Wikipedia articles from the established language modeling benchmark WikiText-103 (Merity et al., 2016), which contains a subset of high-quality Wikipedia articles. This gives us a dataset of 23,522 graph-text pairs in total, covering 82.3% of Wikitext-103 articles. On average each graph has 38.7 nodes and 48.3 edges, and each text article contains 3,533.8 tokens. In addition to structural information, our graphs also contain rich text information with an average of 895.1 tokens in each graph. Furthermore, the automatic process we used to create this dataset can be extended to pair any Wikipedia document with Freebase, and can be scaled up to create over 3M graph-text pairs.

Out of many exciting new tasks that this dataset enables, we present 3 possibilities: graph $\rightarrow$ text generation, graph $\rightarrow$ text retrieval, and text $\rightarrow$ graph retrieval. We benchmarked a few baseline models on these tasks. The models we considered were based on the recent Transformer-XL (Dai et al., 2019) model, and we adapted it to condition the text generation on the KG in different ways. Our results show that better conditioning on the graph indeed improves the relevance of the generated text and the retrieval quality. However, there is still significant room for improvement on these tasks, which makes this an exciting dataset for research. Our data and code for baseline models will be made publicly available.

## 2 Related work

**Graph-text paired data** There has been a lot of prior work on creating graph-text paired datasets. Example applications include generating text summaries conditioned on Abstract Meaning Representation graphs (Liu et al., 2018), generating the abstract of a scientific article given a KG and title (Koncel-Kedziorski et al., 2019) and generating text from RDF triples (Gardent et al., 2017; Jin et al., 2020). Therefore previous KG-text paired datasets that rely on human annotation have limited scale. Among these, Gardent et al. (2017) crowdsourced human annotators to verbalize RDF triplets taken from DBpedia (Auer et al., 2007) to a few sentences (WebNLG) and this caused errors in annotation that were fixed with a few updates through years. Parikh et al. (2020) paired Wikipedia Table with one sentence text that is created by annotators that revise Wikipedia text.

Another line of research focuses on eliminating the need of human annotations by automatically matching KG-text pairs or generating KGs from text using existing tools. Lebret et al. (2016) automatically matched Wikipedia infobox of biographies with their first sentence. Koncel-Kedziorski et al. (2019) utilized an earlier information extraction system that extracts entities, co-reference and relations from given text to build KG’s. The GenWiki dataset (Jin et al., 2020) is automatically constructed by querying KGs in DBpedia with the title of articles in Wikipedia followed by filtering and entity annotation.

We construct our WikiGraphs dataset by extracting a subgraph from Freebase (Bollacker et al., 2008) for each Wikipedia article following a scalable automatic process. Compared to previous work, our WikiGraphs dataset contains significantly larger graphs and longer text (Table 1).

**Models for graph-text paired data** Recent state-of-art language models are based on the Transformer architecture (Vaswani et al., 2017) that uses the self attention mechanism. The Transformer-XL (Dai et al., 2019) model further introduces a segment level recurrence with a novel positional encoding resulting in impressive performance in long sequences by capturing dependencies beyond a fixed length window.

| Dataset     | #examples | #triples | #tokens | #vocab |
|-------------|-----------|----------|---------|--------|
| WebNLG      | 13,036    | 2.54     | 15.26   | 1,484  |
| GenWiki     | 1.3M      | 1.95     | 21.46   | 476,341|
| Ours        | 23,522    | 48.3     | 3,533.8 | 238,071|

Table 1: Our dataset contains significantly larger graphs (average #triples per graph) and longer text (average #tokens per text) than previous KG-text datasets.
the graph structures, making them a good fit for graph data. GNNs have been used in many applications on KG's (Kipf and Welling, 2016; Wang et al., 2019; Xu et al., 2019). Fundamentally, transformers can also be understood as a special type of GNNs with a fully-connected graph structure.

The most recent prior work on graph-to-text generation follows an encoder-decoder architecture (Koncel-Kedziorski et al., 2019; Jin et al., 2020), where the graph part is encoded with a GNN model, e.g. Graph Attention Network (GAT) (Veličković et al., 2018). The text part is typically modeled using an attention based decoder with a copy mechanism (e.g. BiLSTMs as in (Jin et al., 2020)) to process input from both the KG and text.

The models we benchmarked for graph-to-text generation were based on the Transformer-XL architecture and conditioned on the graph through a GNN, making full use of the graph structure and capable of generating very long text comparable to the state-of-the-art.

3 Dataset

In this section we first present some properties of our dataset, and then describe the process that we used to create it.

3.1 Properties of the data

3.1.1 Scale of the data

Basic statistics about our WikiGraphs dataset are listed in Table 2. An illustration of a graph-text pair is shown in Figure 1. A few actual examples from our dataset are included in the Appendix (Figure 7, 8). All of the articles come from the WikiText-103 dataset (Merity et al., 2016), which contains high-quality articles that fit the Good or Featured criteria specified by the Wikipedia editors when the data was collected. Merity et al. (2016) have already cleaned up and tokenized the articles, therefore they appear as plain text without any markup tags.

As will be described in Section 3.2, we try to pair each article with a subgraph from Freebase, centered at the entity node that has a Wikipedia link to the title of the article. We are not able to match every article to an entity in Freebase, but through this process we retained a significant portion of 82.3% of the WikiText-103 articles. We kept the original train/valid/test split. As we will see in Section 4.2, training models on this set gives us results that are very close to training on the full WikiText-103 dataset when evaluated on our test set. Therefore the text part of WikiGraphs appears to be sufficient to reproduce and benchmark against the state-of-the-art text generative models.

Figure 2 shows the distribution of graph sizes and article lengths across our dataset. All the distributions are skewed with a long tail. Notably, average graph size in our dataset is 38.7 nodes and 48.3 edges, considerably larger than the graphs in previous datasets (Jin et al., 2020; Gardent et al., 2017). Also the length of the text articles averages to 3,533.8 tokens and can go up to 26,994 tokens, which is orders of magnitudes longer than the text data in previous graph-text paired datasets that typically only contains a single or few sentences (Jin et al., 2020; Gardent et al., 2017; Lebret et al., 2016).

3.1.2 Nodes and edges

The graphs in our dataset contains two types of nodes: entities and string literals. Each entity is labeled by a unique Freebase entity ID, e.g. ns/m.0f9q9z, and each string literal contains some natural language text, that could be for example a name, date, or description of an entity. Each edge in the graphs also has an associated edge label, e.g. ns/common.topic.description, indicating which type of edge it is. There are a total of 522 different edge types in our dataset. Figure 3 shows the frequency of all the different edge types in our dataset.

Every graph always has one entity node (we call it “center node”) that has a link to the paired Wikipedia article, through a special edge key.wikipedia.en, and the whole graph is a 1-hop neighborhood of entities around the center node within the bigger Freebase KG, plus the string literals associated with all the entities included. Note that it is possible to have edges between the 1-hop neighbors of the center node, therefore the graphs typically are not star structured. Section 3.2
The distribution of information is not uniform across the nodes in a graph. Figure 5 shows that most entity nodes in our graph has a small degree, while few nodes have much larger degrees. Also most string literal nodes contain short text, while fewer nodes contain longer text.

The skewed distribution of nodes and edges in our dataset reflect the nature of KG’s like Freebase, and presents new challenges to graph representation learning models.

3.2 The dataset construction process

We follow three principles when designing the dataset construction process:

1. The text part of the data should be directly comparable in complexity to the capability of state-of-the-art text generative models.
2. The graph part of the data should be constructed in an automatic and scalable way.
3. The graph part of the data should be relevant for the paired text data.

Note that our process is general, and can be applied to any set of Wikipedia articles. We have tried to pair a full dump of English Wikipedia with Freebase and managed to get over 3 million graph-text pairs. Here we restrict the process to the set of articles from the WikiText-103 dataset.

We try to map each Wikipedia article to a relevant subgraph of the existing large scale KG Freebase (Bollacker et al., 2008). We used the last public dump of Freebase\(^2\), which contains 1.9B triples and a total of 250GB of data. We filtered the data by keeping only the entities with at least 4 string attributes (otherwise the entities are less interpretable), and keeping only the top 1024 most frequent relation types and restricting the relations to

\(^2\)https://developers.google.com/freebase
only those among the retained entities and between the entities and string attributes. We also simplified the entity and relation names by stripping off the irrelevant “http://rdf.freebase.com/” and further removed duplicates. This gives us a significantly cleaner and smaller backbone graph for Freebase, with about 20M nodes.

Finding the relevant subgraph for an article in such a cleaned up but still large KG remains non-trivial. Our process for this contains 3 stages: mapping, expansion, and filtering.

**Mapping** In the first stage of the process, we map each article into an entity in our processed Freebase KG. This is made possible through triples from Freebase like the following:

\[
\text{ns/g.11b6jbqpt4 key/wikipedia.en "Madunnella"}
\]

where `ns/g.11b6jbqpt4` refers to an entity in the KG, `key/wikipedia.en` is the type of the edge, which indicates that this entity is linked to a Wikipedia article and “Madunnella” is the title of that article. We normalize the title string (and in general any string literals) from Freebase by replacing “_” with white space and handle unicode characters properly. We extract the titles from the Wikipedia article through string matching, where titles are enclosed in a “= [title] =” pattern.

In this step we managed to map 24,345 out of 28,475 (85.5 %) article titles from WikiText-103 to an entity in our KG.

**Expansion** We treat each of the mapped entities as the center node of a subgraph, and expand 1 hop out in the entire filtered Freebase graph to include all the neighboring entities that are the most relevant to the center entity. We then expand further from this 1-hop graph out to include all the relations that connect the selected entities to string attributes as well as between these entities themselves. Note that because of these edges between the 1-hop neighbor entities the graphs are typically not star structured. This gives us a relevant but compact graph for each article. We have also investigated the possibility of a 2-hop neighborhood from the center node, and found that 2-hop neighborhoods are significantly larger than 1-hop and through some “hub” nodes like “Male” or “Female” a 2-hop neighborhood from an entity can easily include many other irrelevant entities. Based on such observations we decided to use the 1-hop neighborhood to keep the relevance of the subgraph high.

**Filtering** The last stage of the process involves more filtering and cleaning up of the data. We noticed that in Freebase it is common for one entity to have multiple relations of the same type pointing to different string attributes, like the following:

\[
\begin{align*}
\text{ns/m.07c72 key/wikipedia.en "The SImpsons"} \\
\text{ns/m.07c72 key/wikipedia.en "The Simpson"} \\
\text{ns/m.07c72 key/wikipedia.en "The simsons"} \\
\text{ns/m.07c72 key/wikipedia.en "Thr Simpsons"} \\
\text{ns/m.07c72 key/wikipedia.en "The Simpson’s"}
\end{align*}
\]

It is clear that there is a lot of redundancy in this data. We reduced all such edges (from the same entity with the same edge type to string attributes) to a single edge by picking the most “canonical” one. This was done by fitting a unigram model to the characters in the collection of strings and using that model to pick the most likely string.

We also filtered the graphs based on size and created three versions of the data with maximum graph size capped at 256, 512, and 1024 nodes, respectively. All the statistics and results in the rest of the paper are based on graphs with a maximum size of 256, but all versions of the data are made available online.

### 4 Experiments

We perform a set of experiments to showcase how the text and graph information can be combined in a language model. Specifically, we consider three
tasks: text generation conditioned on the graph, graph retrieval given the text, and text retrieval given the graph.

4.1 Graph-conditioned Transformer-XL

In order to incorporate graph information into an advanced language model, we adapt the recent Transformer-XL model (Dai et al., 2019) to also attend to the graph features. At a high-level our model embeds the graph into a set of embedding vectors, and then exposes these embeddings to the Transformer-XL model as extra "token" embeddings to condition on. The size of this set depends on the graph model we choose.

Given the features for $T$ text tokens $H_t \in \mathbb{R}^{T \times d}$ and features for $T'$ graph “tokens” $H_g \in \mathbb{R}^{T' \times d'}$, we illustrate the graph-conditioned attention procedure with a single head as follows:

$$
\begin{align*}
Q_t, K_t, V_t &= H_t W^q_t, H_t W^k_t, H_t W^v_t \\
K_g, V_g &= H_g W^q_g, H_g W^k_g \\
A_t, A_g &= Q_t K_t^T, Q_g K_g^T \\
A, V &= [A_t \circ A_g], [V_t \circ V_g] \\
O &= \text{Masked-Softmax}(A)V
\end{align*}
$$

where $[a \circ b]$ stands for concatenation on the sequence dimension and thus $A \in \mathbb{R}^{T \times (T+T')}$ and $V \in \mathbb{R}^{(T+T') \times d_h}$, where $d_h$ is the head dimension. In other words, comparing to the original Transformer-XL, our model also computes the attention scores between the text queries $Q_t$ and both the text keys $K_t$ and the graph keys $K_g$. As a result, the attention outputs contain information from both the graph and the text context. Note that this formulation is compatible with an additional memory (Dai et al., 2019) with minimal changes, as it simply adds in an extra set of “tokens” for the model to attend to. We don’t use position encodings for the graph “tokens” as there is no sequential ordering for them.

In this work we consider three different approaches for encoding the graph structure:

- **Bag-of-words (BoW):** we construct a single bag-of-words representation of all the tokens from both the nodes and edges in the graph. Entity IDs and numeric values in the graph are replaced with special tokens `<entity>` and `<number>`. The BoW vector is further projected using a linear layer to a latent space. In this case $T' = 1$.

- **Nodes only (Nodes):** we construct separate BoW representations for each node and project each to an embedding and ignore the edges. In this case $T'$ is equal to the number of nodes in the graph.

- **Graph neural network (GNN):** we embed BoW representations for both nodes and edges and then use a graph neural network (Battaglia et al., 2018) on top of those embeddings to compute a new set of node embeddings. $T'$ is equal to the number of nodes.

The $T'$ graph embeddings from this process are shared across all the time steps for text tokens. This model can be further improved, e.g. by using word embeddings and text summarization techniques, but we leave these for future work.

4.1.1 Implementation details

We reimplement the Transformer-XL model in Jax (Bradbury et al., 2018). In our experiments, we employ the base model in (Dai et al., 2019), except that we increase the tail shrinkage factor used for the adaptive softmax and input representations from 1 to 4, which saves 63% of the parameters without compromising the performance. On the full Wikitext-103 dataset, our implementation has a test perplexity of 24.2 (published result for this base model was 24.0). We train our models using the standard likelihood objective for language models with a total batch size of 64 on 8 V100 GPUs. Adam optimizer is used with an initial learning rate of $2.5 \times 10^{-4}$, which decays up to 200k steps following a cosine curve. During training, we use text segments of 150 steps and a memory of equal size. When evaluating the model, we use a sequence length of 64 and memory size 640. Unless further noted, in our experiments we use an embedding size of 256 for BoW-conditioned models. For other models, we project each node or edge represented by BoW to an embedding space of size 128. The default GNN we use has a single linear message passing layer of 256 hidden units.

4.2 Graph $\rightarrow$ text generation

Our first task is text generation conditioned on the graph. We evaluate model performance by (1) computing model perplexity on held-out text and (2) drawing samples from the model and comparing that to the ground truth text article. We use BLEU score (Papineni et al., 2002) to measure the similarity of our generated samples to the ground truth.
Table 3: The perplexity and the generated text reverse-BLEU score of different types of graph-conditioned models. We show the reverse-BLEU score with or without prompting the original title at the start of the text generation.

Unlike previous use cases for BLEU score where there are many references for one generated sample, here we have only one ground truth reference but we can generate multiple samples. We therefore simply swapped the reference with the samples when computing the score, which we term as the reverse-BLEU (rBLEU). We have also tried other ways of computing the BLEU score and find that they don’t change how models compare against each other.

Unless explicitly stated, we let the model sample with a memory size of 640, and condition on the graphs in the test set to generate text for up to 512 tokens per sample for a total of 20 samples per graph. The rBLEU score is computed based on these samples and corresponding ground-truth texts are truncated to the same length. We sample the texts from the distribution with a temperature of 0.8. For each case, we report the average rBLEU score of 3 sampling runs. We find the variances are insignificant which do not affect the comparison results. In Appendix A.3 we also report results for generating longer samples for up to 4096 tokens.

4.2.1 Main result

In Table 3, we show the perplexity and the rBLEU score of the unconditional, BoW, nodes-only, and GNN conditioned models. As a reference, a standard Transformer-XL model trained on the full Wikitext-103 training set reaches 25.08 perplexity on our test set, which contains 71.7% of the original test articles. We can see that the unconditional, i.e. text only, model trained on our dataset gets a very similar performance as trained on the full set. This is strong evidence that our dataset can be a good benchmark for state-of-the-art text generative models.

We also see that conditioned on the graphs, model perplexity didn’t improve, but the relevance of the samples measured by the BLEU scores did improve significantly. This indicates that the graph conditioned models can indeed steer the language model towards more relevant topics, but this so far cannot yet improve likelihood metrics.

To make the evaluation more fair to the text-only model, we also tried to prompt the generation with the title of the article, such that the text-only model also has some context. In this setting the graph models are still better, showing the importance of modeling the structure.

Lastly, among all the 3 graph model variants, we observe that using a set of embeddings from the nodes model is better than using a single embedding from the BoW model, and fully utilizing the graph structure through the GNN model is consistently better than ignoring the edges as in the nodes model. However the differences among the methods are relatively small. For visualizations of a few graphs in our dataset and the corresponding samples generated based on them please refer to Appendix A.

4.2.2 Ablation studies

We show a few ablations on the graph model and sampling parameters, to provide some insights into the models.
Table 4 shows the effect of varying the number of message passing layers in the GNN. We can observe that there is a big difference between using message passing (≥ 1 layers) or not (0 layers) in terms of rBLEU score, but increasing the number of message passing layers does not change the results significantly. We believe however, that these results can be improved by employing bigger and more powerful graph representation learning models, and potentially use initial node and edge representations better than bag-of-words.

In Figure 6 we show the effect of the graph size on model performance. In this experiment we subsample the nodes in each graph to control for the amount of context the model has access to. It is clear from the results that when we heavily subsample and keep only a small portion of the graphs, the GNN model performs similarly as the simpler BoW model, but GNNs benefit more as we keep more of the graph structure.

### 4.3 Graph → text retrieval

In this task, we evaluate the possibility of retrieving relevant text for a given query graph. We pair all articles with all graphs in the test set, resulting in $43 \times 43 = 1849$ pairs. Then the trained graph-conditioned language models are used to produce the per-token likelihood of each pair, and we use these likelihood scores to rank the text articles for each graph. We expect the learned models can rank the correct pairs higher than wrong ones. To measure the results we use standard ranking metrics including recall@K, which computes the fraction of times the correct pair is included in the top K predictions, as well as mean average precision (mAP). In Table 5, it is observed that graph-conditioned models can indeed retrieve more relevant texts from the graph than the unconditional model, among which the GNN-based model performs the best, and the unconditional model performs close to a random guess.

### 4.4 Text → graph retrieval

In this last task, we evaluate the performance of graph retrieval given a text query. We use exactly the same setting and scores as Section 4.3, but instead rank the graphs for each text article using the likelihood scores. The results are shown in Table 6. Note that this task is quite easy with our data and setup, potentially because the graphs are much more distinguishable than the text articles. All the graph-conditioned models perform almost perfectly, with the GNN model again outperforming the others.

### 5 Conclusion

In this paper, we present WikiGraphs, a new graph-text paired dataset with significantly larger graphs and longer text compared to previous datasets of similar nature. We show that the text part of this data is a good benchmark for state-of-the-art text generation models, and the paired dataset can help us benchmark models that are capable of generating long and coherent text conditioned on a graph structure.

In the first set of experiments on this dataset we showcase 3 different tasks using our dataset, and demonstrate the benefit of better models that make more use of the graph structure.

There is still significant room for improvement for these tasks on our dataset, and we hope the release of the data and baseline code can help spur more interest in developing models that can generate long text conditioned on graphs, and generate graphs given text, which is another exciting direction our dataset enables but we did not explore, and eventually bridging the graph and text modalities.

### References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.
Loïc Barrault, Magdalena Biesalska, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanos, et al. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. JAX: composable transformations of Python+NumPy programs.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

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.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR.

Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. GenWiki: A dataset of 1.3 million content-sharing text and graphs for unsupervised graph-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2398–2409, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. *Text Generation from Knowledge Graphs with Graph Transformers*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. *Microsoft coco: Common objects in context*. In *European conference on computer vision*, pages 740–755. Springer.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A Smith. 2018. Toward abstractive summarization using semantic representations. *arXiv preprint arXiv:1805.10399*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

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.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuvan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset.
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In International Conference on Learning Representations.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 39(4):652–663.

Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. 2019. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 968–977.

Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. 2019. Cross-lingual knowledge graph alignment via graph matching neural network. arXiv preprint arXiv:1905.11605.

A Appendix

A.1 Graph visualization
Some example visualizations of the KG structures are shown in Figure 7 and Figure 8. The corresponding graph truth texts are shown in Table 7.

A.2 Generated examples
The generated texts based on the graph shown in Figure 7 and Figure 8 are listed in Table 8 and Table 9, respectively.

A.3 Ablations on sampling configurations
We show additional ablation results on the sample length (Table 10) and the temperature (Table 11) for greedy sampling. Note that for each case we show the rBLEU score based on the validation set computed with a single sampling run (20 samples per graph).

Note that the GNN model has overall the best performance. However as the sample length increases the advantage of the GNN model also decreases. This indicates that it is still very challenging to generate long text that stays on-topic, and potentially the noise overwhelms the signal when number of tokens increases to 4096.
Figure 7: Visualization of the “Where the Streets Have No Name” KG in our dataset.
Bourbon County is a county located in Southeast Kansas. As of the 2010 census, the county population was 15,173. Its county seat and most populous city is Fort Scott.

Figure 8: Visualization of the “Fort Scott National Historic Site” KG in our dataset.
“Where the Streets Have No Name” is a song by Irish rock band U2. It is the opening track from their 1987 album The Joshua Tree and was released as the album’s third single in August 1987. The song’shook is a repeating guitar arpeggio using a delay effect, played during the song’s introduction and again at the end. Lead vocalist Bono wrote the lyrics in response to the notion that it is possible to identify a person’s religion and income based on the street on which they lived, particularly in Belfast. During the band’s difficulties recording the song, producer Brian Eno considered erasing the song’s tapes to have them start from scratch. “Where the Streets Have No Name” was praised by critics and became a commercial success, peaking at number thirteen in the US, number fourteen in Canada, number ten in the Netherlands, and number four in the United Kingdom. The song has remained a staple of their live act since the song debuted in 1987 on The Joshua Tree Tour. The song was performed on a Los Angeles rooftop for the filming of its music video, which won a Grammy Award for Best Performance Music Video.

Writing and recording

The music for “Where the Streets Have No Name” originated from a demo that guitarist The Edge composed the night before the group resumed The Joshua Tree sessions. In an upstairs room at Melbeach House — his newly purchased home — The Edge used a four-track tape machine to record an arrangement of keyboards, bass, guitar, and a drum machine. Realising that the album sessions were approaching the end and that the band were short on exceptional live songs, The Edge wanted to “conjure up the ultimate U2 live 0-0 song”, so he imagined what he would like to hear at a future U2 show if he were a fan. After finishing the rough mix, he felt he had come up with “the most amazing guitar part and song of [his] life”. With no one in the house to share the demo with, The Edge recalls dancing around and punching the air in celebration. Although the band liked the demo, it was difficult for them to record the song. Bassist Adam Clayton said, “At the time it sounded like a foreign language, whereas now we understand how it works.” The arrangement, with two time signature shifts and frequent chord changes, was rehearsed many times, but the group struggled to get a performance they liked. According to co-0-0 producer Daniel Lanois, “that was the science project song.”

Fort Scott National Historic Site

Fort Scott National Historic Site is a historical area under the control of the United States National Park Service in Bourbon County, Kansas, United States. Named after General Winfield Scott, who achieved renown during the Mexican–American War, during the middle of the 19th century the fort served as a military base for US Army action in what was the edge of settlement in 1850. For the next quarter century, it was used as a supply base and to provide security in turbulent areas during the opening of the West to settlement, a period which included Bleeding Kansas and the American Civil War. The current national historic site protects 20 historic structures, a parade ground, and five acres (20,000 m²) of restored <unk> prairie, inside the city of Fort Scott. It is open to visitors most days of the year.

History

In 1842, Fort Scott was named after Winfield Scott, who was established on the American frontier on the military road in eastern Kansas between Fort Leavenworth and Fort Gibson. It was established to provide protection to the rapidly increasing number of settlers, who were migrating from the Eastern United States. Fort Scott became one of a chain of forts intended to protect the new settlers from the Plains Indians, as well as to protect the Indians from the settlers’ encroachment. The United States government intention to reserve permanent Indian lands west of the Missouri River gave way to the competition of settlers continuing to encroach on the Indian settlements. Fort Scott’s most active days were between 1842 and 1853, although it was also used during the Civil War.

Army days

The Cherokee of Indian Territory (now Oklahoma) were upset to have Fort Wayne in their proximity. After some delay, the US Army decided to abandon Fort Wayne and move its soldiers to a new fort to be built between Fort Leavenworth and the site. The Army both wanted to placate the Cherokees (who were supervised by the Department of Defense) and provide more defense for white settlers and other Indians against the Osage, who had been conducting frequent raids in the area. On April 1, 1842, some soldiers of Fort Wayne left their fort and on April 22 arrived where Fort Scott would be built, in the Osage <unk> section of modern 0-0 day Kansas. After bickering with the Cherokees to acquire the land, the rest of Fort Wayne’s garrison left the fort on May 26 and arrived at the Fort Scott site on May 30. Unlike most forts for military use, the fort did not have defensive walls or structures.

Table 7: Ground truth samples.
You Can’t Take This Away (U2 song)

You Can’t Take This Away (U2 song) is a song by Irish rock band U2, released as a single. It was written by vocalist Bono and produced by The Smiths for their third solo album, The Joshua Tree. Inspired by Romania roots rock and roll, “You Can’t Take This Away” is a song about a man who suffers from severe night terrors. The song was first released on the album and premiered on radio on 19 August 1987, and it was included on the band’s Billboard Hot 100. “You Can’t Take This Away (U2 song)” received mixed reviews from critics; some tracks, such as the opening chorus and bridge, were praised as highlights by some reviewers. The song was well received by critics, as the record company’s highest selling single at that time, where “The Edge of Forever” peaked at number 64. It was later re-released on the band’s 2006 compilation album No Line on the Horizon, but has since been re-released on live performances in 2006 and 2009. “You Can’t Take This Away (U2 song)” was performed on the Late Show with David Letterman on 31 December 2005. “You Can’t Take This Away (U2 song)” has since been covered by many groups and has achieved enormous commercial success. A video for the song was filmed by the @-@frontman Bono, for which it was nominated for a Grammy Award.

The Edge of Glory

The Edge of Glory is a song by Irish rock band U2, released as a single. It was written by guitarist Larry Mullen, Jr., who plays lead guitar on the song, and was produced by Alex Cink, who described the song as “a track with a lot of meaning, but no connection.” The song contains several pop rock elements and is set in that time period, and is among the most prominent in the album. In addition to its lyrics, the song’s lyrics detail hypnotherapy, and also deals with the effects of adultery. The song’s lyrics have been described by music critics as being autobiographical. The lyrics have been described as “a bold exploration of the figure of a New York City man,” and “an expression of the inability of freedom to live in a world that is also a place in the world of space.” The song’s lyrics describe a “Manhattan @-@ like place,” with Bono calling the arrival a “pleasant little optimism from before it came to life.” The Edge of Glory was a success in the United Kingdom, reaching number two in the charts in the United States, and topping the charts in Australia and New Zealand. The song has been certified platinum by the Recording Industry Association of America, and has sold over four million copies worldwide. The song has been covered by several artists, including German band U2. The music video for “The Edge of Glory” won Best Video at the 2004 MTV Video Music Awards. The video also served as an inspiration for the film U2 360°.

Numb

“Numb” is a song by U2, released as a single from their album Zoo Station. It was written and produced by Niall Horan and produced by U2 and the Edge. “Numb” received mixed reviews from critics; some tracks, such as the opening chorus and bridge, were praised as highlights by some reviewers. The song was well received by critics, as the record company’s highest selling single at that time, where “The Edge of Forever” peaked at number 64. It was later re-released on the band’s 2006 compilation album No Line on the Horizon, but has since been re-released on live performances in 2006 and 2009. “You Can’t Take This Away (U2 song)” was performed on the Late Show with David Letterman on 31 December 2005. “You Can’t Take This Away (U2 song)” has since been covered by many groups and has achieved enormous commercial success. A video for the song was filmed by the @-@frontman Bono, for which it was nominated for a Grammy Award.

Zoo Station

“Zoo Station” is a song by Irish rock band U2, released as a single. It was written by guitarist Larry Mullen, Jr., who plays lead guitar on the song, and was produced by Alex Cink, who described the song as “a track with a lot of meaning, but no connection.” The song contains several pop rock elements and is set in that time period, and is among the most prominent in the album. In addition to its lyrics, the song’s lyrics detail hypnotherapy, and also deals with the effects of adultery. The song’s lyrics have been described by music critics as being autobiographical. The lyrics have been described as “a bold exploration of the figure of a New York City man,” and “an expression of the inability of freedom to live in a world that is also a place in the world of space.” The song’s lyrics describe a “Manhattan @-@ like place,” with Bono calling the arrival a “pleasant little optimism from before it came to life.” The Edge of Glory was a success in the United Kingdom, reaching number two in the charts in the United States, and topping the charts in Australia and New Zealand. The song has been certified platinum by the Recording Industry Association of America, and has sold over four million copies worldwide. The song has been covered by several artists, including German band U2. The music video for “The Edge of Glory” won Best Video at the 2004 MTV Video Music Awards. The video also served as an inspiration for the film U2 360°.

Node

“Zoo Station” is a song by Australian rock band U2 from their second studio album, PopMart, released in 1992 as the album’s lead single. It was recorded and produced by the Edge, and was released on 20 August 1992. The song was written by Larry Mullen, Jr., the song’s lead singer, and was produced by Alex Cink, who described the song as “a track with a lot of meaning, but no connection.” The song contains several pop rock elements and is set in that time period, and is among the most prominent in the album. In addition to its lyrics, the song’s lyrics detail hypnotherapy, and also deals with the effects of adultery. The song’s lyrics have been described by music critics as being autobiographical. The lyrics have been described as “a bold exploration of the figure of a New York City man,” and “an expression of the inability of freedom to live in a world that is also a place in the world of space.” The song’s lyrics describe a “Manhattan @-@ like place,” with Bono calling the arrival a “pleasant little optimism from before it came to life.” The Edge of Glory was a success in the United Kingdom, reaching number two in the charts in the United States, and topping the charts in Australia and New Zealand. The song has been certified platinum by the Recording Industry Association of America, and has sold over four million copies worldwide. The song has been covered by several artists, including German band U2. The music video for “The Edge of Glory” won Best Video at the 2004 MTV Video Music Awards. The video also served as an inspiration for the film U2 360°.

Background

“Zoo Station” was written by U2 frontman Bono and produced by U2 and was U2’s first single on the album. Clayton still had a strong fanbase and was a regular singer. The two met while performing in the amphitheaters of Los Angeles, and in a live performance he was the lead singer on the last leg of the Vertigo Tour. The Edge and Clayton were both fans of the band, and the pair decided to collaborate on the album. Both performed on their 2004 tour, and made a solo appearance on the 2004 The Zoo TV Tour. Clayton and Clayton had been close friends, and the pair became friends again in 2004.

Table 8: Generated samples based on the “Where the Streets Have No Name” graph.
Cond. Generated Text

= Carcurel =

Bloodlines is a 2011 action role @-@ @ of playing video game developed by <unk> and published by Activision for the Wii. The game is part of a sequel of the 2013 game, The Wolf Among Us, and was released on November 19, 2011. It was published by Activision for Xbox 360, Xbox 360, and Xbox 360. The game, set in the fictional Middle Ages, revolves around the player @-@ character centuries battling for control of their universe. The protagonist of the game, Percy, the leader of the Knights of the Old Republic, is a member of the Knights of the Old Republic, and is appointed to lead a military coup to overthrow the Irish Republic and destroy the Romé Nations' military forces. Though the game mainly took place in a new version of the New York City, the original plan was to make it more easily accessible to players unfamiliar with the New Republic. It was also a commercial success, selling more than 900,000 copies. The game received mostly positive reviews from most video game publications, with many praising the visual style and the gameplay, but many said that it was not as good as that of the previous game. Reviewers noted the game's slight forward addressing issues such as the difficulty level, a general danger of being too difficult to fight, and the difficulty of playing the game as the player @-@ character's pattern of characters.

= Gameplay =

Bloodlines is a crossover action role @-@ @ of playing game that takes place in the fictional Middle Ages, which is composed of medieval countries and localities. Valhalla, a medieval stronghold, is the game's main setting. The player @-@ character is a 3 @-@ 3D miniature character with a sword and shield, which have multiple colored attacks, and has two of the four abilities, which are progressively reduced from the first one and allow for greater size and movement. The available weapons are bolt @-@ fired weapons, advanced weapons, and weapons that can be used in battle. The player is able to summon magical powers to attack targets, and can use magical powers to enhance the character's abilities. <unk> are also available via a <unk> system, which enables players to throw stones at enemies and attack enemy characters who have not encountered them. The player character also has an ability to revive foes by performing a touch @-@ touch action. The game can be played as a side @-@ @ of scrolling through a View Mode, which can be used in the single @-@ @ player mode. The first act features a <unk> displayed from a first @-@ @ person perspective. The player character can move around.

BoW

= Civil War Pass =

Civil War Pass, also known as the Battle of the Creek or the Battle of Fort Sumner, was an important battle fought on September 7, 1864, at Fort Coldwater, in the state of Montana. After seeing repeated attacks on the fort, Gen. James A. Douglas, the commander of the Confederate forces in the South, decided to abandon the fort and flee to the north. After Union forces struck the fort, they decided to flee south to the Ohio River. There they quickly encountered a group of horses, who were used to build a pontoon bridge. The ditches and wooden planks were removed and replaced with stone blocks to make them float (place). The obstacles that were created in the river valley, however, proved treacherous and were not bridged by mountain passes. The young general and his troops eventually reached the Ohio and the Mississippi rivers, but the new Presidential candidate, Abraham Lincoln, resigned after the war. After the defeat at Fort Sumner, General Douglas, the commander of the Union forces, planned and executed a number of attacks on Fort Sumner. When soldiers arrived, they found two now @-@ @ deserted locations. The attacks had been made more than a year before, and when the line of retreat of the Union forces, which stretched from Fort Sumner to Fort Duxbury, was completed on August 19, 1864, the cavalrymen captured it on September 30. In November 1864, General Douglas was defeated at the Battle of Lake Logan.

= Background =

In 1861, with the Mexican @-@ American War nearing its conclusion, the American public began to think of an armistice treaty, or peace treaty between Mexico and the United States. On July 1, 1861, General Douglas sent three large armies from the Mexican @-@ American War, a series of forts west of the Rockies, to renew attacks on the Union. On September 17, 1861, the British Forces abandoned the fort, and the American Army was allowed to capture the fort. Confederate forces launched a cavalry attack on Fort Dix in early 1865, but both sides suffered defeats. The British and Americans were unprepared for the chance of victory, and the Americans were now planning to take control of the Gulf Coast. As the Americans, the British planning an attack into central Canada. The British were aware that the main American forces would be found in the south of Canada, so they occupied the town of York. The British were near the Niagara River and the Americans were hopefully moving along the river, approaching Fort Sumner from the west. The British were reluctant to move toward the Carolinas, and so, in the event the Port of Boston was abandoned, the British would be forced to travel to the lower Mississippi. The

Nodes

= Fort Scott =

Fort Scott is an American military post located in Fort Lee, Kansas. It is named in honor of General William Scott, a U.S. Army general and the first commander of the Army of the Potomac. The site was designated as a National Historic Landmark in 1991, and has been designated a National Historic Landmark under the title of Fort Scott Historical Site since 1929. It is located in the Rocky Mountains in Kansas and is known as the "James Scott National Historic Site."

= History =

The original fort was built on the Caddo on the Black River, and later moved to Fort Lee in present @-@ @ day Decatur County, Virginia. On July 10, 1810, the government of the United States established a military post at Fort Scott. The fort was constructed on a site that had previously been occupied by a small group of people. Around 1800, the first Governor of Kansas, Colonel Andrew H. Sharpe, established Fort Scott in what is now the state of Montana. The fort was named after an American Indian, and was known as Fort Douglas. The fort was used for administrative purposes and for the administration of the Territory of Missouri. In 1808, William Bolivar Buckner led a major invasion to remove the western boundary of Texas, which included Fort Scott. Congress authorized a survey of the territory in 1817, and a survey of the Old South boundary was completed in 1818, making Fort Scott the first governor to apply federal law. Although the West Texas Act of 1845 quickly became a national concern, the new governor was unable to raise sufficient funds to maintain Fort Scott. The fort's construction and completion were completed in 1821, but the state legislature refused to grant the commissioners the land they were granted. The new land, which was purchased by the U.S. Army, was named after General John Scott, a U.S. Army general and the first governor of Texas, who was killed in an assassination attempt on June 20, 1855. military historian John P. Pickett wrote that it was the first military governor of the United States in the treaty of Fort Scott.

GNN

= Fort Scott National Historical Park =

Fort Scott National Historical Park is an Illinois state historic park in the U.S. state of Kentucky. It is located at 53°20′44″ N 85°41′16″ W, at the edge of the Clay Creek valley, southwest of New Orleans. It is located at the intersection of Washington Boulevard and State Route 63, and is the largest National Historic Landmark in the state. The park was purchased by Native Americans in 1903, and the site was added to the National Register of Historic Places in 1962. Since 1998, the site has been subject to an extensive series of historic markers and features that are important in preservation of American historic sites in Texas. The National Park Service includes the nation's oldest extant log cabins, historic buildings, historic markers, and historic sites. The park is home to the Mississippi River National Historical Park, a U.S. National Monument that supplies historic sites and historic sites. The original fort was built in 1818 to protect U.S. statehood. In 1889, the state legislature constructed a small blockhouse at the site of the original fort to defend it from Native Americans. The blockhouse first appeared in 1868, when land in the city of Lisbon was granted to the state. The fort has remained in use since then.

= History =

= = Early history =

Fort Scott was established as a civil and military fortress in 1803 and named after an American Indian, the land that would become Fort Scott was originally part of the Louisiana Purchase, which was granted to the United States by the Louisiana Purchase Act of 1825. The original fort was established in 1828 by an act of Congress. The American Revolutionary War came to an end in 1830, but Independence was declared in 1831 and Independence was declared on June 3, 1830. The post @-@ war Treaty of Paris signed at Fort Scott ended military activity in the region. War by the United States reached an end in 1830, and most of the land was put aside for use as a military park. Fort Scott was garrisoned by 350 soldiers from the 5th Louisiana Regiment during the War of 1812. In 1837, the Illinois General Assembly passed a bill to establish the Fort Scott as a federal park in the same year the state agreed to purchase the site in honor of the site's new state of Louisiana. Originally, only about half of Fort Scott was owned, but the size of the park changed in the 1880s from a forest reserve to a desert road. The park was significantly expanded during the 1910s, but the exact date is disputed.

Table 9: Generated samples based on the "Fort Scott National Historic Site" graph.
| Cond. | Valid rBLEU (w/ title) |
|-------|------------------------|
| Sample length | | |
| 256   | 512 | 1024 | 2048 | 4096 |
| None  | 9.53 | 10.47 | 12.22 | 14.57 | 14.60 |
| BoW   | 30.63 | 29.44 | 29.56 | 29.92 | 30.00 |
| Nodes | 32.33 | 30.30 | 29.82 | 30.43 | 29.91 |
| GNN   | 33.81 | 31.32 | 30.39 | 30.53 | 30.05 |

| Cond. | Valid rBLEU (w/ title) |
|-------|------------------------|
| Sample length | | |
| 256   | 512 | 1024 | 2048 | 4096 |
| None  | 12.08 | 10.47 | 9.71 | 27.09 | 27.78 | 26.21 |
| BoW   | 28.21 | 29.44 | 27.63 | 31.25 | 32.48 | 31.02 |
| Nodes | 29.55 | 30.30 | 28.48 | 31.52 | 32.88 | 31.23 |
| GNN   | 29.59 | 31.32 | 29.01 | 31.55 | 32.49 | 31.20 |

Table 10: Generated samples vs sample length.

Table 11: Generated samples vs temperature.