GenWiki: A Dataset of 1.3 Million Content-Sharing Text and Graphs for Unsupervised Graph-to-Text Generation

Zhijing Jin*
Amazon Shanghai AI Lab
zhijing.jin@connect.hku.hk

Qipeng Guo*†
Fudan University
qpguo16@fudan.edu.cn

Xipeng Qiu
Fudan University
xpqiu@fudan.edu.cn

Zheng Zhang
Amazon Shanghai AI Lab
zhaz@amazon.com

Abstract

Data collection for the knowledge graph-to-text generation is expensive. As a result, research on unsupervised models has emerged as an active field recently. However, most unsupervised models have to use non-parallel versions of existing small supervised datasets, which largely constrain their potential. In this paper, we propose a large-scale, general-domain dataset, GenWiki. Our unsupervised dataset has 1.3M text and graph examples, respectively. With a human-annotated test set, we provide this new benchmark dataset for future research on unsupervised text generation from knowledge graphs.[1]

1 Introduction

Text generation with deep learning models is data hungry. For example, in Figure 1, to make a model learn how to verbalize knowledge graphs, researchers need to collect a large number of human-annotated text and graph pairs. However, good annotation is both expensive and difficult to get – annotators need to have a thorough understanding of hundreds of edge types in the knowledge graphs, as well as proper verbalization of the text, so that the written text can conform to the distribution of the desired text style. Moreover, for dataset curators, checking the quality of annotation is also non-trivial. For example, the WebNLG dataset goes through five updates to fix errors in the annotation over the past 3 years.[2]

These difficulties in dataset collection makes parallel data-to-text datasets small-sized, and even non-existent for low-resource domains. To make problems worse, data-to-text models are, in many cases, infeasible to transfer from one domain to another. For example, a text generation model that can produce Wikipedia biography-like descriptions cannot be used to generate introductions of plants. This can happen even between domains with similar content but different text styles. For example, given the knowledge graph triple "(Obama, birthYear, 1961)," one domain verbalizes it as "Obama was born in 1961," whereas another domain prefers "Obama (1961 –) ...". A model trained in the first domain can only generate the entities correctly but fail on all other words in the second domain.

To overcome the lack of labelled data and difficulty in domain adaptation, unsupervised data-to-text generation has emerged as an active research field recently (Freitag and Roy, 2018; Schmitt et al., 2020; Guo et al., 2020). However, the progress of this line of research is slowed down due to the lack of large-scale unsupervised datasets. Notably, the curation of graph-to-text unsupervised datasets are non-trivial, as it requires (1) same content distribution between graphs and text, (2) text with high-accuracy entity annotation, (3) a much larger scale than the supervised datasets, and (4) a human-annotated test set. Unfortunately, lacking such an unsupervised dataset, most unsupervised works have to artificially remove the pairing information between text and structured data, to force parallel datasets to be non-parallel. Obviously, splitting parallel datasets, such as the WebNLG dataset (13K), and E2E dataset (50K), into non-parallel ones remains the originally small data size. Consequenty, the research on unsupervised

*Equal contribution.

†Work done during internship at Amazon Shanghai AI Lab.

[1]Our dataset is available at https://github.com/zhijing-jin/genwiki.

[2]https://gitlab.com/shimorina/webnlg-dataset

Proceedings of the 28th International Conference on Computational Linguistics, pages 2398–2409
Barcelona, Spain (Online), December 8-13, 2020

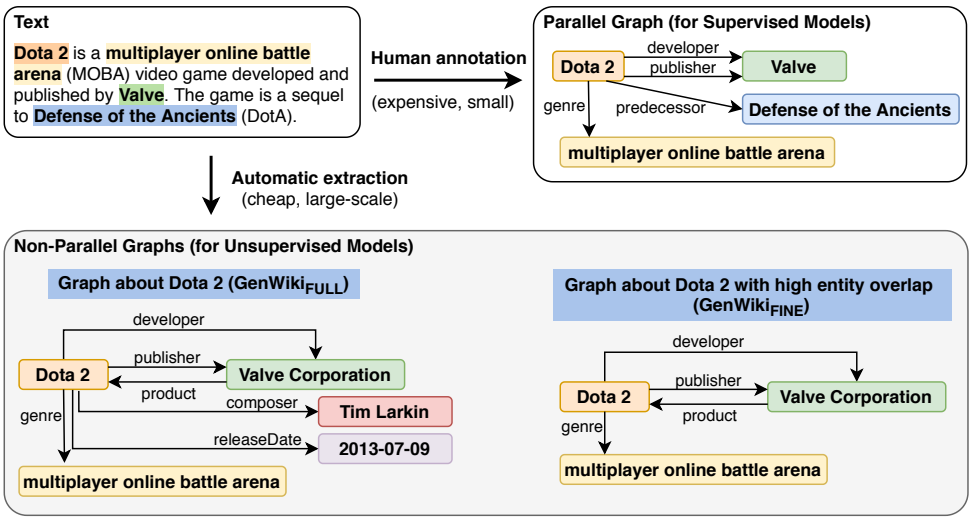


Figure 1: Overview of the dataset.

models is limited, as these existing unsupervised models cannot even have a deep architecture. In contrast, a relatively faster field, unsupervised machine translation, has dataset sizes on the order of billions, such as 1.6B German and 2.1B English text used in (Artetxe et al., 2019).

Therefore, we propose a large dataset, GenWiki, which contains 1.3 million non-parallel text and graphs with shared content, and meet all four requirements mentioned before. To better facilitate research in unsupervised graph-to-text generation, we provide two versions of our dataset: the full dataset GenWiki$_{\text{FULL}}$ (1.3M), and a fine version, GenWiki$_{\text{FINE}}$ (750K), which adds constraints on the text and graphs to force them to contain highly overlapped entity sets. The overview of our two datasets are shown in Figure 1. The GenWiki$_{\text{FULL}}$ on the bottom left contains graphs on the same topic (Dota 2) as the text, and GenWiki$_{\text{FINE}}$ imposes a stronger constraint that entities in the graph can largely overlap with entities in text. Both datasets are collected in a scalable, automatic way. The comparison of our dataset and previous data-to-text datasets is illustrated in Table 1.

An additional contribution is our human-annotated test set with of 1,000 graph and text pairs. Based on our large-scale training dataset and the human-annotated test set, we analyze the performance of several baselines and existing models. We conducted error analysis on the strengths and shortness of these unsupervised models in Section 5.4.

Dataset	Size	Purpose	Collection	Domain
KBGen (Banik et al., 2013)	0.2K	Supervised	Human	Descriptions of biology knowledge bases
RoboCup (Chen and Mooney, 2008)	0.7K	Supervised	Human	Sportscast
RotoWire (Wiseman et al., 2017)	5K	Supervised	Human	Baseketball game summaries
SF Hotels/Restaurants (Wen et al., 2015)	10K	Supervised	Human	Dialogues about restaurants and hotels
WebNLG (Gardent et al., 2017)	13K	Supervised	Human	15 categories (building, person)
WeatherGov (Liang et al., 2009)	22K	Supervised	Human	Weather forecasts
E2E (Novikova et al., 2017)	50K	Supervised	Human	Restaurant and hotel descriptions
WikiCompany (Qader et al., 2018)	51K	Supervised	Human	Company descriptions
ToTTO (Parikh et al., 2020)	100K	Supervised	Human	Description of Wikipedia tables
WikiBio (Lebret et al., 2016)	500K[3]	Supervised	Auto	First sentence of Wikipedia biographies
GenWiki$_{\text{FINE}}$	750K	Unsupervised	Auto&distant align	General domain (all wiki topics)
GenWiki$_{\text{FULL}}$	1.3M	Unsupervised	Auto	General domain (all wiki topics)

Table 1: Comparison of previous text generation datasets and our GenWiki datasets. GenWiki$_{\text{FINE}}$ is the fine version (with stronger entity alignment) of our dataset, and GenWiki$_{\text{FULL}}$ is our full dataset.

[3]The unfiltered dataset contained 728K, but only 500K are non-empty samples.

2 Existing Data-to-Text Models and Datasets

Research in NLP can be viewed as a close interplay between models and datasets. Traditionally, most methods hand-craft linguistic features to build rule-based systems, so datasets containing hundreds or thousands of samples are adequate. For example, in data-to-text generation, the traditional approach is to hand-craft templates (Kukich, 1983; Holmes-Higgin, 1994; McRoy et al., 2000). There are also works that abstract the templates and use probabilistic models to learn the rules (Angeli et al., 2010; Howald et al., 2013). Popular datasets for these methods are, for example, the 0.7K RoboCup dataset (Chen and Mooney, 2008).

With the advance of deep learning, models have a large number of parameters, so they have to be fueled by large datasets. For data-to-text generation, many supervised methods use large neural networks. To match with these data-hungry models, researchers devote many efforts to curate supervised datasets. Most supervised data-to-text datasets have tens of thousands of data.[4] For instance, WebNLG dataset (Gardent et al., 2017) has 13K valid data-text pairs,[5] WeatherGov (Liang et al., 2009) has 22K weather forecasts, and E2E (Novikova et al., 2017) has 50K restaurant and hotel descriptions, as shown in Table 1.

Recently, there is a rising trend of unsupervised approaches (Freitag and Roy, 2018; Schmitt et al., 2020; Guo et al., 2020). Unsupervised data-to-text models (Konstas and Lapata, 2012) are proposed in response to the lack of data-text pairs in many domains, similar to the emergence of unsupervised machine translation that addresses lack of data low-resource language pairs. However, to match with the active research on unsupervised data-to-text generation, *no* suitable dataset has been curated. As a result, most previous works have to artificially force parallel datasets to be non-parallel, by separately shuffling the data part and text part of the original dataset, or directly deleting the text part. However, such formulation will make the unsupervised datasets inherit the small size of the supervised datasets, which are limited to only tens of thousands samples.

There are several caveats of using small datasets for unsupervised models. (1) Limiting the model potential: Unsupervised models can be data-hungry. For example, 1.6B German and 2.1B English tokens are fed to unsupervised machine translation models (Artetxe et al., 2019); 3.3 billion words are used to pretrain the language model BERT (Devlin et al., 2019). With abundant data, the potential of unsupervised models is unleashed – impressive performance, based on an enormous number of parameters, such as 12 layers of transformers. So far, no large data-to-text corpus can be used to unveil the potential of unsupervised data-to-text. (2) Lack of diversity: As many unsupervised models need to impose strong priors such as grammar and dependency tree (Konstas and Lapata, 2012), if a proposed model works well on a specific type of text generation, we cannot validate whether such a model generalizes well. (3) Negligence of model efficiency: The goal of unsupervised models is to learn from as many unsupervised data as possible and thus achieve high accuracy. However, the current designs of many unsupervised models do not take efficiency into consideration because of the small datasets they use does not expose the problem of efficiency.

Hence, to foster research in unsupervised data-to-text learning, we construct a new, large-scale dataset, GenWiki, which can satisfy all constraints imposed by existing unsupervised models. The dataset will be introduced in Section 3 and 4.

3 Desiderata

Although it is easy to collect unsupervised datasets for tasks such as machine translation, it is a different story for data-to-text generation. To match with the design of unsupervised data-to-text models, some strict constraints are required. A common constraint is that the text and knowledge graphs should have the same content distribution. Some more specific constraints require, for example, the text corpus to have entity annotations. The reason is that recent unsupervised learning models (Guo et al., 2020; Schmitt et al., 2020) use cycle training of two tasks: graph-to-text, and text-to-graph, which is simplified to relation extraction *given entities*. This simplification requires the unsupervised text corpus to have

[4] https://aclweb.org/aclwiki/Data_sets_for_NLG

[5] (Moryossef et al., 2019) report that there are 13K clean data-text pairs, despite the original 25K data.

entity annotations in text. Additionally, it is also reasonable to make the relation types in the knowledge graphs a closed set of, for example, 300 relations as in WebNLG (Gardent et al., 2017). In this way, the relation extraction model can avoid collapsing on unseen relations in the test set, and it is also an easier job for human annotators as they only need to consider a limited-sized, well-defined set of relations.

We summarize the requirements of the dataset collection as follows:

- Basic Requirements
  1. Text and graphs should have similar contents, such as a Wikipedia article and a knowledge graph of the same article.
  2. To fit for recent unsupervised models (Guo et al., 2020; Schmitt et al., 2020), the text corpus should contain entity annotations.
  3. The knowledge graphs should have a closed set of relations.
- Preferred Properties
  4. Large scale (over 500K).
  5. Diversity, not limited to one specialized domain.
  6. Human-annotated test set, as opposed to distant supervision test sets (Mintz et al., 2009; Riedel et al., 2010), because noisy test sets can misguide the comparison of unsupervised models.

4 GenWiki Dataset

Based on the desiderata outlined in Section 3, we will first introduce the construction process of our training set in Section 4.1, and test set in Section 4.2. We will then analyze the characteristics of the resulting dataset in Section 4.3.

4.1 Training Set Construction

An ideal unsupervised graph-to-text dataset allows text sequences consisting of multiple sentences, and graphs of several triples. In the collection process of GenWiki, we allow each text sequence to have 1 to 10 sentences and at most 50 words, and graphs to have 1 to 10 triples. Our dataset aims to satisfy all the requirements mentioned in Section 3, namely (1) same content distribution between text and graph, (2) entity annotation in text, (3) a closed set of relations, (4) a large size, (5) diversity, and (6) a human-annotated test set. Specifically, our construction process is as follows.

Step 1. Data Collection Our dataset, GenWiki, is collected from general Wikipedia. Different from other data that are limited to specific categories of Wikipedia (Gardent et al., 2017; Lebret et al., 2016; Wang et al., 2018), we aim at general-domain data-to-text generation. To this end, we collect text from the HTML webpages of all Wikipedia articles by April 2020, and use the titles of these articles to query their corresponding knowledge graphs from DBpedia.[6] Note that we identify all hyperlinked terms in the HTML files as entities. We tokenize the text corpus using the NLTK package.[7]

The output of this step is Wikipedia articles with hyperlinked entities, and each article's corresponding knowledge graph.

Step 2. Filtering For all the articles retrieved from Wikipedia, we filter out empty pages, and pages containing placeholder contents such as "See the error message at the bottom of this page for more information." To control the quality, we also look at the top 10 frequent sentences, and filter out high-frequency but unrelated sentences such as "Media related to ... at Wikimedia Commons."

For the noisy graphs collected from Step 1, we only keep relations (namely edge types) that can be grounded to DBpedia relation ontology. We then filter out meta-relations such as "wikiPageRedirect," and relations that are unlikely to be described in text such as "latitude" and "longitude." After this filtering, due to the close-set relation constraint, we only keep the top 300 frequent relations, similar to the practice of (Gardent et al., 2017). The full list of relations that are filtered out or kept is available at `https://github.com/zhijing-jin/genwiki`.

[6] `http://dbpedia.org/isparql/`

[7] We use the NLTK package (`http://nltk.org/`) only for experimentation.

After Step 2, the resulting dataset include 14,734,778 articles, and 3,009,112 graphs, with 19.45 triples per graph on average.

Step 3. Entity Annotation To fit for the recent advances of unsupervised models based on cycle training, we take a challenging step to annotate the entities in text. From Step 1, terms with hyperlinks are annotated as entities. However, such annotation is very sparse, one per 14.50 words. The reason is that Wikipedia hyperlinks are manually added by human contributors, and not all occurrences of entities are linked. For example, there is no hyperlink in the sentence "In October 1871, Claude Monet returned to France." despite the existence of three entities, "October 1871," "Claude Monet," and "France."

To increase the entity annotation in text, we develop a hybrid of methods based on the following intuitions:

1. All entities in the graph should be annotated if they also occur in the corresponding article.
2. The surface forms of graph entities (e.g., "President Obama" is a surface form of "Barack Obama") should also be annotated if they appear in the corresponding article.
3. Named entities, including dates, locations, organizations, and numbers, should be annotated.
4. Personal pronouns (e.g., "he", "she") should also be annotated.

Therefore, we first prepare a set of candidate entities for each sentence, including (1) entities in the knowledge graph corresponding to the Wikipedia article, (2) all surface forms of graph entities, (3) named entities annotated by the stanza package,[8] and (4) personal pronouns except the versatile "it." We annotate these entities whenever their occurrence in the text is detected. Note that there might be overlapping entities (e.g., "Obama" is a substring of "Barack Obama," but both are entities), so we start from entities with the most number of words to entities with the fewest number of words.

The resulting density of annotated entities in text is one per 4.28 words.

Step 4. Text-Graph Alignment Remember the first hard requirement (that text and graphs should have similar contents). We introduce an alignment step to ensure that this constraint can be approximated. This step is inspired by the assumption introduced in distant supervision for relation extraction (Mintz et al., 2009; Riedel et al., 2010): if the entity sets of a text sequence and a relation overlap, it is highly likely that the sentence express that relation.

In Step 4, for each article and its corresponding graph, we enumerate all text sequences (1-10 subsequent sentences with at most 50 words) and all small graphs (1-10 triples). For each text sequence, we only add it to the text corpus if its entity set can overlap with the entity set. Similarly, for each small graph, we add it to the graph dataset only if its entity set overlaps with the entity set of at least one text sequence. In this way, we collect the text sequences and graphs with shared content, and form $\text{GenWiki}_{\text{FULL}}$.

It is also good to curate a finer version of our dataset by imposing a higher threshold of entity overlap rate. Specifically, we construct $\text{GenWiki}_{\text{FINE}}$ which consists of text sequences that have an F1 score of over 40% and entity overlap of at least 2 with at least one graph, and graphs in the same way.

4.2 Test Set Annotation

Apart from the unsupervised training set, we also collect a human-annotated test set of 1,000 text-graph pairs. To ensure the quality, we recruit annotators who have experience in data-to-text research. The annotators are introduced the background of the task, format of the data, and goals of annotation.

As the test set annotation is intensive and challenging, we design a workflow to minimize the annotators' effort. Based on a randomly selected sample set from the training set, we automatically construct distantly aligned graph-text pairs, to serve as a reference for human annotators. Each distantly aligned pair are the text and graph which have a higher entity F1 with each other than with any other text or graph of the same Wikipedia article. The whole workflow is as follows:

1. We construct distantly aligned pairs, and, as supplementary information, we show the annotator the entire knowledge graph (up to a hundred triples) of the source Wikipedia article of each pair.

[8] https://stanfordnlp.github.io/stanza/

2. We format each example in a reader-friendly way, by underlining all entities in text, and bold-facing overlapped entities between text and graphs. We also provide the corresponding Wikipedia article title, so that the annotator can look up the article if the annotation requires background knowledge.
3. Annotators first check if there is an easy way to edit the text and graph so that they express the same content.
    - If yes, then make minimal edits to align the text and graph
    - If no, then split one sample into two cases, where the first case keeps the text as original and edits the graph to align with the text, and the second case keeps the graph as original and modifies the text.
4. We use programs to check whether in each annotated sample, the text entities perfectly match with the graph entities. If not, we reject the sample, and return it to the annotators to redo Step 3.

4.3 Analysis

Reflection on the Desiderata We compare our dataset with the two most similar previous datasets, WikiBio (Lebret et al., 2016) and WebNLG (Gardent et al., 2017). The properties evaluated in Table 2 correspond to the desiderata we outlined previously in Section 3. Unlisted properties imply that all three datasets have satisfied them. The diversity property is evaluated by both the domain and text type.

	Basic Requirements	**Preferred Properties**			
	Entity Annotation	>500K	Domain	Text Type	Gold Test
WikiBio	✗	✓	Biography	1st sentence of biography	✗
WebNLG	✓	✗	15 specific categories	1-7 crowdsourced sentences	✓
GenWiki	✓	✓	All wiki categories	1-10 Wikipedia sentences	✓

Table 2: Properties of WikiBio, WebNLG, and our dataset with regard to the desiderata in Section 3.

Among the three datasets, WikiBio, despite a large size, does not have entity annotations or a gold test set, so it cannot be used by many unsupervised models. On the other hand, WebNLG is a human-annotated supervised dataset, so it has high-quality entity annotation, and covers relatively diverse topics, 13 wiki categories including athletes, buildings, and universities. However, due to the manual annotation, WebNLG cannot scale to higher orders of magnitudes such as millions of data samples that our dataset has.

Our dataset, GenWiki, satisfies all criteria. Its text corpus has entity annotation, and GenWiki has a gold test. Due to the automatic collection process, it has a large number of data, and can easily scale to other domains in the future. It is also diverse in terms of categories of Wikipedia articles that every data sample is extracted from. The diversity can also be reflected through the text type, as crowd-sourced sentences tend to be simpler and have less variations than the well-edited Wikipedia text. We will quantify characteristics of our dataset compared to WebNLG more in detail later in this section.

Dataset Overview For our two versions of datasets, GenWiki$_{\text{FINE}}$ and GenWiki$_{\text{FULL}}$, we summarize the overall statistics of our GenWiki dataset in Table 3. We compare it with WebNLG, which also meets all three basic requirements, and has been used for previous unsupervised models (Guo et al., 2020; Schmitt et al., 2020).

	Overview		**Graph**			**Text**		
	Examples	Entities	Triples	Avg Triples	Relations	Tokens	Vocab	Avg Len
WebNLG	13,036	1,727	33,075	2.54±1.42	244	198,927	1,484	15.26±8.13
GenWiki$_{\text{FINE}}$	757,152	1,230,920	2,000,636	2.64±1.72	287	19,725,390	328,487	26.05±10.99
GenWiki$_{\text{FULL}}$	1,336,766	1,950,664	2,607,997	1.95±1.42	290	28,693,319	476,341	21.46±10.64

Table 3: Overall statistics of GenWiki$_{\text{FINE}}$ and GenWiki$_{\text{FULL}}$, compared with WebNLG.

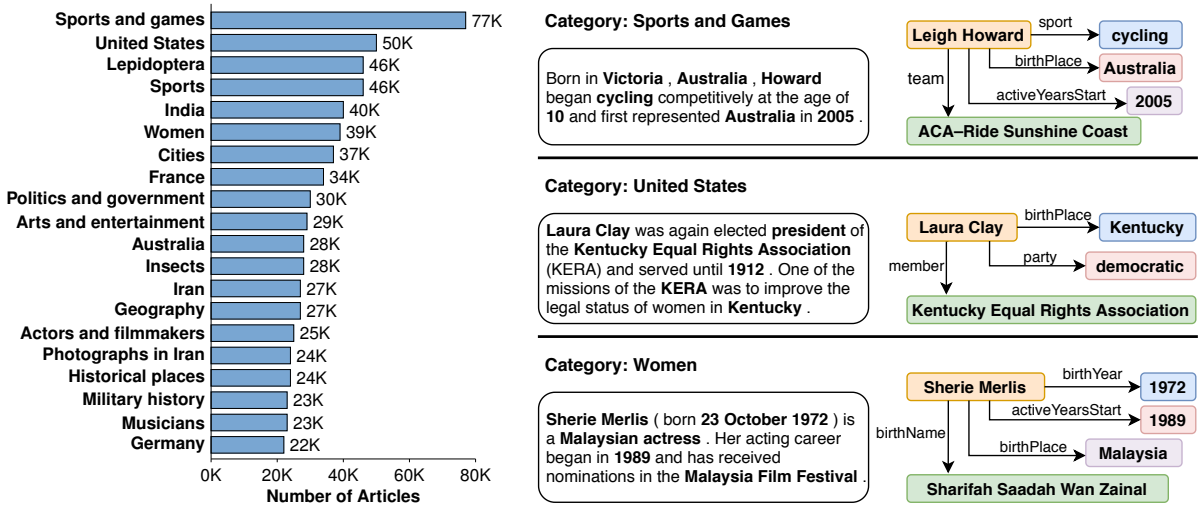
We can see from Table 3 that our dataset has significantly more data than the human-annotated WebNLG, and can be a better dataset for unsupervised learning. Our GenWiki$_{\text{FINE}}$ contains 757K ex-

amples with about 20M tokens, and GenWiki$_{\text{FULL}}$ contains 1.3M samples with about 30M tokens. The number of examples of our GenWiki$_{\text{FINE}}$ is 58 times the size of WebNLG, and GenWiki$_{\text{FULL}}$ is 102 times the size of WebNLG. Similar to the overall size, the number of entities, triples, tokens, and vocabulary size of our dataset are also several orders of magnitudes higher than that of WebNLG. On average, each graph of GenWiki$_{\text{FINE}}$ has 2.64 triples and each text 26.05 words; GenWiki$_{\text{FULL}}$ has 1.94 triples per graph, and 21.46 words per text sequence. The smaller average size of GenWiki$_{\text{FULL}}$ than GenWiki$_{\text{FINE}}$ might be due to the lower entity overlap threshold when collecting data for GenWiki$_{\text{FULL}}$, which allows for more small graphs, and short text sequences.

Diversity of Topics We plot the most frequent topics and their counts in Figure 2a, and show some examples of three representative categories in Figure 2b. To make Figure 2a more illustrative, we omit the top-1 frequent category, biography, which has 339K Wikipedia articles, and plot the next 20 categories which are more on the same scale.

Figure 2a shows that the dataset includes a variety of topic categories such as sports and games (77K), politics and government (30K), and arts and entertainment (29K). It also include many places, ranging from United States (50K), to India (40K), and France (34K). It also has nature-related articles, such as the order of insects Lepidoptera (46K), insects (28K), and geography (27K). Interestingly, there is a considerable number of articles related to women (39K).

Figure 2b shows some typical examples in three categories, sports and games, United States, and women. Bold terms in the text box are automatically annotated entities. Note that in the dataset, each topic has some related graphs and text. Although not strictly aligned, the graphs and text have an overall similar content distribution.



(a) Most frequent categories of topics and corresponding counts in GenWiki$_{\text{FULL}}$.

(b) Non-parallel text and graphs of three frequent categories, sports and games, United States, and women.

Figure 2: Top categories in GenWiki and examples.

Distribution of Relation Types Relation types is an important feature as it indicates the diversity of knowledge graphs, and correlates with the diversity of text generation. As we can see from Figure 3, there are a variety of relations, from the biography-related ones such as birthPlace, birthYear, and deathPlace, to general relations such as country, family, activeYearsStart, populationTotal, isPartOf, and so on. The distribution of relations are long-tailed, as some least frequent relations, although not plotted, have only 10 to 20 occurrences.

Richness of Text and Graphs We evaluate the lexical richness and graph characteristics in Table 4. We first show the lexical diversity by type-token ratio (TTR), on which WebNLG scores 0.007, compared to 0.016 and 0.017 of the two GenWiki datasets. As TTR is sensitive to sample size, we also show the mean segmental type-token ratio (MSTTR) (Johnson, 1944). MSTTR is the average TTR for successive

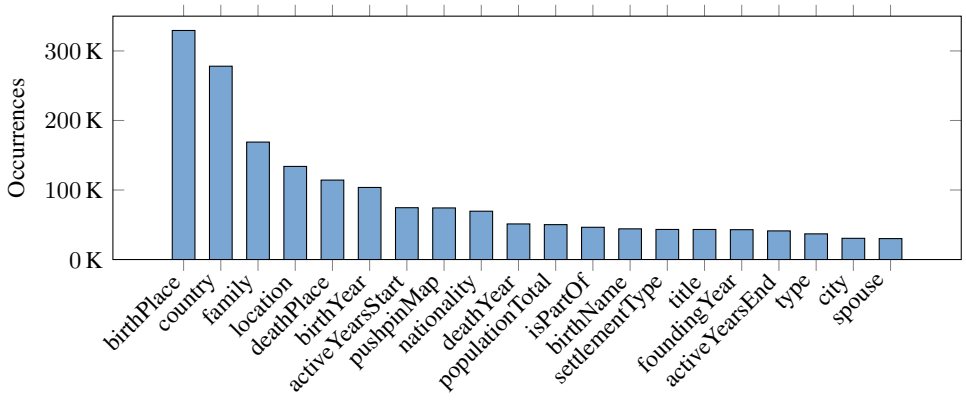


Figure 3: Top 20 most frequent relations and corresponding counts on GenWiki$_{\text{FULL}}$.

segments of text or transcript that contain a standard number of tokens. We calculate the MSTTR for successive every 50 words in the text corpus of each dataset. WebNLG's MSTTR is lower than that of GenWik by a large margin. We also evaluate the entity density (the number of entity words divided by the number of all words), where we can see that all three datasets have a similar entity density of around 23% or 27%. Complementary to the entity density, we also define the concept of information density, which refers to the number of words divided by the number of triples, because text and graphs share the same content. Apart from characteristics of text richness, we also look into the graph richness, and calculate the percentage of graphs with more than two triples in Table 4.

	TTR	MSTTR	Entity Words in Text	Information Density	Graphs with $\geq$2 Triples
WebNLG	0.007	0.46	23%	6.01	69%
GenWiki$_{\text{FINE}}$	0.016	0.72	27%	9.86	72%
GenWiki$_{\text{FULL}}$	0.017	0.72	23%	11.00	48%

Table 4: Dataset characteristics. Information Density is number of words per triple.

5 Evaluating Unsupervised Data-to-Text Models

Since the main purpose of our GenWiki dataset is to serve for future models on unsupervised graph-to-text generation, we provide a preliminary set of experiments using previous unsupervised data-to-text models. Our results can serve as baselines for future work developed on this resource.

5.1 Models

Based on our datasets, we evaluate the following unsupervised baselines.

Rule-Based As a baseline proposed by (Schmitt et al., 2020), Rule-Based linearizes the graph into a sequence of triples. Each triple is described by turning the camel-cased relation type to a normal phrase. For example, the triple "(New York City, populationTotal, 8 million)" will be verbalized as "New York City population total 8 million." The descriptions of multiple triples are joined by "and."

DirectTransfer DirectTransfer is another intuitive baseline, where we use a model trained on the supervised WebNLG dataset, and test it on the GenWiki test set. For a fair comparison, we use a graph transformer generation model proposed by (Koncel-Kedziorski et al., 2019) for all graph-to-text models.

NoisySupervised We also propose another baseline, NoisySupervised, which attempts to absorb the weak supervision signals in the training set, and convert the difficult unsupervised learning problem into supervised learning. Specifically, NoisySupervised first constructs distantly aligned pairs on the whole training data, using the same matching method that we adopted to create our preliminary test set before human annotation. It then takes these pairs with noises as supervision signals, and learn from them using the graph transformer (Koncel-Kedziorski et al., 2019).

CycleGT$_{\text{Base}}$ (Guo et al., 2020) proposes two models, the first of which uses a basic setting, iterative back translation with no pretraining. The CycleGT model jointly learns from two losses, a graph-to-text generation loss, and text-to-graph relation classification loss. Such a cycle training setting, despite its simplicity, was proved comparable performance with supervised models on the WebNLG dataset.

CycleGT$_{\text{Warm}}$ The second setting proposed in (Guo et al., 2020) uses a warm up strategy before the cycle training process. Specifically, as entities are given by the dataset and serve as shared information between text and graphs, the CycleGT model warms up by learning entity-to-text generation and entity-to-graph relation classification. This warm up strategy is used to make the unsupervised training more stable.

5.2 Evaluation Metrics

We evaluate the text generation quality by four commonly used metrics: BLEU (Papineni et al., 2002), Meteor (Banerjee and Lavie, 2005), ROUGE$_{\text{L}}$ (Lin, 2004) and CIDEr (Vedantam et al., 2015), to measure the closeness of the reconstructed paragraph (model output) to the input paragraph.[9]

5.3 Implementation Details

For CycleGT$_{\text{Base}}$ and CycleGT$_{\text{Warm}}$, we use the same code provided by (Guo et al., 2020). For DirectTransfer and NoisySupervised, we use the graph transformer model (Koncel-Kedziorski et al., 2019) reimplemented by (Guo et al., 2020) using the Deep Graph Library (DGL) (Wang et al., 2019). For a fair comparison, we use the same hyperparameters for the overlapped components of NoisySupervised, CycleGT$_{\text{Base}}$, and CycleGT$_{\text{Warm}}$.

5.4 Results

	GenWiki$_{\text{FINE}}$				GenWiki$_{\text{FULL}}$			
	BLEU	METEOR	ROUGE$_{\text{L}}$	CIDEr	BLEU	METEOR	ROUGE$_{\text{L}}$	CIDEr
Rule-Based	13.45	30.72	40.93	1.26	13.45	30.72	40.93	1.26
DirectTransfer	13.89	25.76	39.75	1.26	13.89	25.76	39.75	1.26
NoisySupervised	30.12	28.12	56.96	2.52	35.03	33.45	58.14	2.63
CycleGT$_{\text{Warm}}$	41.35	35.20	63.01	3.45	40.47	34.84	63.40	3.48
CycleGT$_{\text{Base}}$	41.59	35.72	63.31	3.57	41.29	35.39	63.73	3.53

Table 5: The performance of unsupervised models on GenWiki$_{\text{FINE}}$ and GenWiki$_{\text{FULL}}$.

Main Results The main experiment results are listed in Table 5. From the text generation quality of the five models, we can see that the Rule-Based baseline (Schmitt et al., 2020) performs poorly on our GenWiki dataset. The DirectTransfer also has a similar performance as Rule-Based, which implies that even though WebNLG and GenWiki are similar wiki-based datasets, it is difficult to make a model trained on one corpus to do well on the other slightly different one. The NoisySupervised model performs relatively well, scoring 30.30 BLEU points. As it relies on the quality of distant supervision on our non-parallel training set, the good performance of the NoisySupervised model validates that the GenWiki dataset has a good alignment of text and graphs. The two models proposed by (Guo et al., 2020), CycleGT$_{\text{Base}}$ and CycleGT$_{\text{Warm}}$ is the strongest out of all unsupervised models, outperforming NoisySupervised by 10 BLEU points. The two models achieve similar performance with each other, with CycleGT$_{\text{Warm}}$ slightly better on BLEU, and CycleGT$_{\text{Base}}$ higher on the other three metrics.

Error Analysis We select 100 samples from the test set, and analyze the outputs of the best performing model, CycleGT$_{\text{Base}}$. There are two main types of error: The first type is that the generated text misses some information in the graph, which comprises 27% of the errors. For example, the ground truth sentence is "Martial Outlaw is written by Thomas Ritz and Pierre David, produced by Pierre David, and directed by Kurt Anderson.", whereas the model misses a lot of information in its output "Martial Outlaw is a song written by Kurt Anderson and Pierre David." The second error type is that the generated text lacks commonsense, which comprises 20% of the errors. For example, the ground truth is "Solar

[9]We calculate all metrics using the pycocoevalcap tool (`https://github.com/salaniz/pycocoevalcap`).

Entertainment Corporation is founded and owned by the Tieng family." but the model generates "The Tieng family was developed by Solar Entertainment Corporation." which violates common sense.

6 Conclusion

In this paper, we constructed a large-scale, general-domain text generation dataset, GenWiki. This work is designed for the emerging research on unsupervised models for text generation. Our dataset is several magnitudes larger than previous text generation datasets, covers a diverse range of topics, and contains more complicated linguistic structures. It relieves the unsupervised models of the limited sizes of previous datasets, and lays the foundation for more future research on unsupervised text generation models.

Acknowledgements

We thank all lab mates at AWS Shanghai AI lab for fruitful discussions and suggestions. We also appreciate the reviewers for their helpful and constructive inputs that help us improve this paper.

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 502–512. ACL.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 194–203. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Eva Banik, Claire Gardent, and Eric Kow. 2013. The kbgen challenge. In Albert Gatt and Horacio Saggion, editors, *ENLG 2013 - Proceedings of the 14th European Workshop on Natural Language Generation, August 8-9, 2013, Sofia, Bulgaria*, pages 94–97. The Association for Computer Linguistics.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 128–135. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Markus Freitag and Scott Roy. 2018. Unsupervised natural language generation with denoising autoencoders. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3922–3929. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from RDF data. In José M. Alonso, Alberto Bugarín, and Ehud Reiter, editors, *Proceedings of the 10th International Conference on Natural Language Generation, INLG 2017, Santiago de Compostela, Spain, September 4-7, 2017*, pages 124–133. Association for Computational Linguistics.

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020. Cyclegt: Unsupervised graph-to-text and text-to-graph generation via cycle training. *CoRR*, abs/2006.04702.

Paul Holmes-Higgin. 1994. *Text generation - using discourse strategies and focus constraints to generate natural language text* by kathleen r. mckeown, cambridge university press, 1992, pp 246, £13.95, ISBN 0-521-43802-0. *Knowledge Eng. Review*, 9(4):421–422.

Blake Howald, Ravikumar Kondadadi, and Frank Schilder. 2013. Domain adaptable semantic clustering in statistical NLG. In Katrin Erk and Alexander Koller, editors, *Proceedings of the 10th International Conference on Computational Semantics, IWCS 2013, March 19-22, 2013, University of Potsdam, Potsdam, Germany*, pages 143–154. The Association for Computer Linguistics.

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2284–2293. Association for Computational Linguistics.

Ioannis Konstas and Mirella Lapata. 2012. Unsupervised concept-to-text generation with hypergraphs. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 752–761. The Association for Computational Linguistics.

Karen Kukich. 1983. Design of a knowledge-based report generator. In Mitchell P. Marcus, editor, *21st Annual Meeting of the Association for Computational Linguistics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, June 15-17, 1983*, pages 145–150. ACL.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1203–1213. The Association for Computational Linguistics.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In Keh-Yih Su, Jian Su, and Janyce Wiebe, editors, *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 91–99. The Association for Computer Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Susan W McRoy, Songsak Channarukul, and Syed S Ali. 2000. Yag: A template-based generator for real-time systems. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 264–267.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, and Janyce Wiebe, editors, *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011. The Association for Computer Linguistics.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2267–2277. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis, editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 201–206. Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. 2018. Generation of company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 254–263. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In José L. Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer.

Martin Schmitt, Sahand Sharifzadeh, Volker Tresp, and Hinrich Schütze. 2020. An unsupervised joint system for text generation from knowledge graphs and semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, November. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Qingyun Wang, Xiaoman Pan, Lifu Huang, Boliang Zhang, Zhiying Jiang, Heng Ji, and Kevin Knight. 2018. Describing a knowledge base. In Emiel Krahmer, Albert Gatt, and Martijn Goudbeek, editors, *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 10–21. Association for Computational Linguistics.

Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander J. Smola, and Zheng Zhang. 2019. Deep graph library: Towards efficient and scalable deep learning on graphs. *CoRR*, abs/1909.01315.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1711–1721. The Association for Computational Linguistics.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2253–2263. Association for Computational Linguistics.