Domain-independent Data-to-Text Generation for Open Data

Andreas Burgdorf, Micaela Barkmann, André Pomp and Tobias Meisen
Chair of Technologies and Management of Digital Transformation, University of Wuppertal, Wuppertal, Germany

Keywords: Open Data, Data to Text Generation, Natural Language Generation, Transformer, Semantic Data Management.

Abstract: As a result of the efforts of the Open Data movements, the number of Open Data portals and the amount of data published in them is steadily increasing. An aspect that increases the utilizability of data enormously but is nevertheless often neglected is the enrichment of data with textual data documentation. However, the creation of descriptions of sufficient quality is time-consuming and thus cost-intensive. One approach to solving this problem is Data to text generation which creates descriptions to raw data. In the past, promising results were achieved on data from Wikipedia. Based on a seq2seq model developed for such purposes, we investigate whether this technique can also be applied in the Open Data domain and the associated challenges. In three studies, we reproduce the results obtained from a previous work and apply them to additional datasets with new challenges in terms of data nature and data volume. We can conclude that previous methods are not suitable to be applied in the Open Data sector without further modification, but the results still exceed our expectations and show the potential of applicability.

1 INTRODUCTION

Nowadays, large amounts of heterogeneous data are produced daily in various contexts. In order to promote public development based on data, the *Open Data Charter* (ODC) (ODC, 2013) was initiated. The ODC is an international collaboration between experts and governments with the aim of appropriately making data publicly available. Five principles have been elaborated for the implementation of the Charter and the achievement of its goals: (1) *Open Data by Default*, (2) *Quality and Quantity*, (3) *Useable by All*, (4) *Releasing Data for Improved Governance* and (5) *Releasing Data for Innovation*. Thousands of Open Data Portals have been created, often with a regional or thematic focus. To generate an added value and to comply with the five principles, data sets require a good infrastructure to be easily searchable and discoverable. However, Burgdorf et al. (Burgdorf et al., 2020) observed that the ODC does not prescribe any standardizations, and thus unhindered access to data is not guaranteed. *Metadata* provided along data published on Open Data Portals poses a critical aspect for the (re-)usability of the data.

As Chandola and Booker (Chandola and Booker, 2022), we understand metadata as all information provided in addition to the actual data to be published. Metadata are, so to speak, ”data about data” containing information on the origin, type, interpretation, dates, descriptions, etc. They specify the given data and provide context. Metadata can help make the associated data easier to understand and interpret by data consumers. Unfortunately, the lack of standardization, synonyms, various formats(Schauppenlehner and Muhar, 2018), ambiguity (Tygel et al., 2016) or the absence of metadata hinder reuse. This circumstance becomes even more pronounced when more data portals have to be visited to aggregate certain data (Burgdorf et al., 2020). Nevertheless, metadata is an important source of information, both for the processing systems that eventually add semantic meanings to data and the human consumer. For both purposes, textual and human readable data documentation is a crucial part of metadata. Burgdorf et al. propose to build exactly this bridge between modern *Natural Language Processing* methods and *Semantic Modeling*. That being said, they outlined different research directions to achieve this objective, such as the identification of methods and the compilation of a data set aligned with the requirements. Nevertheless, the proposed research perspective depends on the availability of textual data documentation for all collected data. However, these are not always available, human-readable, or of good quality (Schauppenlehner and Muhar, 2018). Furthermore, the authors argue that, if provided, there is also no guarantee that the

Burgdorf, A., Barkmann, M., Pomp, A. and Meisen, T.
Domain-independent Data-to-Text Generation for Open Data.
DOI: 10.5220/0011272900003269
In *Proceedings of the 11th International Conference on Data Science, Technology and Applications (DATA 2022)*, pages 95-106
ISBN: 978-989-758-583-8; ISSN: 2184-285X

textual data documentation actually correlates with the associated data.

To enrich the data landscape of Open Data Portals on the one hand and to support research in metadata-driven automated *Semantic Data Management* on the other hand, we propose an automated and domain-independent generation approach for textual data documentation under the usage of *Data-To-Text Generation* methods. For this purpose, we investigate whether we can successfully apply existing data-to-text generation models to the open domain sector and whether we can achieve satisfactory results even with minimal amounts of data. We also examine which evaluation method is best suited to assess the quality of generated data documentation.

In the remainder of this study we provide a brief introduction into the Natural Language Generation approach we use for our experiments. We present three different experiments utilizing different data sets and their results and discuss how well the selected method performed in each experiment. Finally, we give an outlook on what has to be done to actually implement NLG in the open data sector.

2 METHODS

To examine the use of Data-to-Text Generation in the Open Data sector, this paper is based on the work of Chen et al. (Chen et al., 2019) who placed similar requirements on a Data-to-Text-Generation (DTG) model as we do except for the application in Open Data Portals.

Chen et al. present a structure-aware seq2seq model based on the GPT-2 language model. It encodes field information from a given table into the cell memory and state of an LSTM. It thus allows an internal structural representation of the given table within the framework. This is achieved using a modified LSTM which has an additional *field gate*. To efficiently incorporate the additional information about the table into the generation, they employ a *Dual Attention* mechanism that allows for both word-level attention and *field-level attention*. Finally, to teach the model when to copy values from the table and when to generate new words, the authors include a trainable function that calculates a *copy probability* for copying versus generation. Above all, they showcase Few-Shot settings with only 50-500 training instances across multiple domains. They achieved great performance and outperformed previous best BLEU baselines by 8.0 points. The methods used provide a good basis for approaching our research objectives. In our study we first tried to replicate the results of Chen et al. (Chen et al., 2019) using the WikiBio data set (Lebret et al., 2016). Beyond that we applied the model to the domain-overlapping ToTTo (Parikh et al., 2020) data set to test its generalization capabilities. In a third step we applied the model to the VC-SLAM (Burgdorf et al., 2022) data set, which is based on data from Open Data Portals and which is very limited in quantity. This allows us to examine how realistic a Few-Show setting is in the real open data sector.

In the following, we will introduce the main ideas of Chen et al. (Chen et al., 2019) and the theoretical setup of their model. We will then present the data sets used in this work whereby we leave out the practical processing steps of these data sets for fitting to the model for now. Finally, we encounter the evaluation methods used in this work.

2.1 Baseline Method

The authors start with the statement that conventional neural-based end-to-end approaches for NLG that take structured data or knowledge as input are very "data-hungry" (Chen et al., 2019, p. 1). As this makes their "adoption for real-world applications difficult" (Chen et al., 2019, p. 1), the authors propose the task of Few-Shot Data-To-Text Generation. With the underlying research questions (1) Can we significantly reduce human annotation effort to achieve reasonable performance using neural NLG models?; and (2) Can we make the best generative pre-training, as prior knowledge, to generate text from structured data?; the authors introduce a model architecture based on content-selection from input data and on generating natural language text with the help of a pre-trained Language Model.

According to the authors, one needs two skills to describe information in a table: (1) select and copy factual content from the table; and (2) compose grammatically correct sentences that bring those facts together, whereby the second skill is not restricted to any domain. The task of forming fluent and coherent sentences can thus be detached from the task-specific components of DTG and be presented in the form of a pre-trained Language Model that represents the "innate" language skill of the neural DTG model. In this way, the authors bypassed data-intensive training because the content-selection skill can be learned "relatively quickly" (Chen et al., 2019, p. 1).

As previously mentioned, the authors used an architecture separated into a content-selection mechanism and a pre-trained Language Model. A switch policy is applied to decouple the framework into those tasks. Figure 1 shows a schematic sketch of the ap-

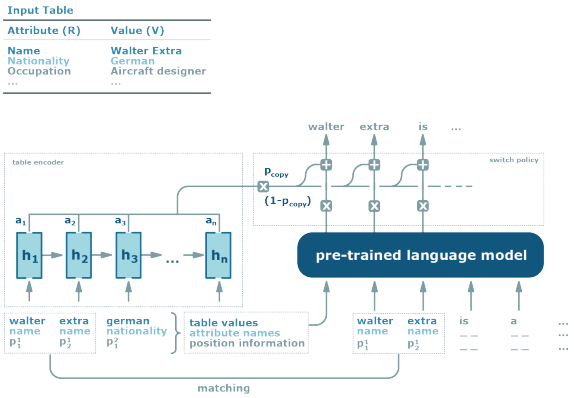

Figure 1: Overview of the approach of Chen et al. (Chen et al., 2019). Illustration modified and adapted from (Chen et al., 2019, p 2).

proach.

Original Problem Formulation. The input data is semi-structured. The goal is to automatically generate a natural language description based on that data, based on only a few hundred training instances. We have semi-structured data in the form of attribute-value pairs, formalized with:

$$\{R_i : V_i\}_{i=1}^{n} \tag{1}$$

With R_i representing the attribute and V_i representing the values and of a table size i. Both, R_i and V_i, can either be a number, a phrase or a sentence. Further, each value is represented as a sequence j of instances:

$$\{R_i : V_i\}_{i=1}^{n} \tag{2}$$

This leads to the effect that for each instance of information v_j, all the information about its attribute R_i and its position in the value sequence is available.

Language Model. For generation, Chen et al. (Chen et al., 2019) used a pre-trained Language Model. The currently most prominent Neural Language Models are *GPT* from OpenAI (GPT-1: (Radford et al., 2018); GPT-2: (Radford et al., 2019); GPT-3: (Brown et al., 2020)) and *BERT* (Devlin et al., 2018). The language model used here is *GPT-2*. It is a transformer-based NLM trained on a data set of 8 million web pages (approximately 40 GB of text). The published GPT-2 model contains 117 million parameters and 12 Transformer layers. In their model, Radford et al. (Radford et al., 2019) chose to represent the input according to *Byte-Pair Encoding* (BPE).

2.2 Data Sets

In this part, the data sets used for the studies are presented in more detail.

WikiBio Data Set. The Wikipedia Biography Data Set[1] (short WikiBio) gathers approximately 728,000 biographies from Wikipedia. Rémi Lebret, David Grangier, and Michael Auli in 2016 (Lebret et al., 2016) built this data set in connection with their work "Neural Text Generation from Structured Data with Application to the Biography Domain". Their paper introduced a neural model for DTG, which generates biographical sentences from fact tables. In contrast to prevailing related works which experimented on DTG, this self-created data set was significantly larger than existing data sets up to that point. Compared to other popular data sets used in the context of DTG, this data set contains around 728,000 samples with a vocabulary of over 400,000 words. Essentially, the data set consists of two parts: a text part and associated structured data.

ToTTo Data Set. The ToTTo data set was published at the beginning of 2020 by Parikh et al. (Parikh et al., 2020). It is a designated open-domain data-to-text data set in the English language with over 120,000 instances. It consists of tables taken from different domains and articles from Wikipedia. In order to prevent overlaps with the WikiBio data set (Lebret et al., 2016), Wikipedia infoboxes were excluded from the collection. When choosing Wikipedia tables, table-sentence pairs were selected that overlapped in at least three non-zero digits. This way, mainly statistical tables were included.

VC-SLAM Data Set. The VC-SLAM (*Versatile Corpus for Semantic Labeling and Modeling*) (Burgdorf et al., 2022) corpus originally comes from a different domain than DTG, namely OBDM. The focus of the corpus is to advance the developmental landscape of *Semantic Mapping* as an essential part of OBDM. In this context, Semantic Mapping describes the process of mapping an attribute (e.g., from a data set) with the corresponding entry in an ontology (Burgdorf et al., 2022). Unlike typical DTG data sets, the corpus contains an ontology along with an associated data set. The data set itself consists of 101 data sets coming from different domains of the *"(smart) city"* context. To implement this limitation in the collection, only data records containing geo-references were considered. However, under the assumption that

[1] https://github.com/DavidGrangier/wikipedia-biography-dataset

the data falls under this context, they can come from many different domains such as speed limits, public restrooms, or air pollution (Burgdorf et al., 2022, p. 7). The data was gathered from an extensive search of over 190 Open Data Portals. In addition to the context criterion, the data also had to provide a textual data description in English to be included in the set.

2.3 Evaluation

In our studies, we use the BLEU-4 (Papineni et al., 2002), ROUGE-4 (Lin, 2004), and PARENT (Dhingra et al., 2019) evaluation methods suitable for the evaulation of DTG results. The BLEU-4 values will be calculated and documented by the model during training so that values will be available over time in intervals of 30 epochs. ROUGE-4 scores are calculated separately and retrospectively. For the PARENT scores, we will set the λ weight to 0.5 for all studies. Whilst BLEU and ROUGE are not optimal evaluation methods for DTG, they have been and still are used in the majority of cases. The usage of the PARENT score can be useful when combined with data sets such as WikiBio, as the textual data documentation and associated tabular data do not offer too much room for interpretation and inference. Therefore, in the potential summary of data records, there is proportionally less variation possible than with, for example, larger historical data sets, ROUGE and BLEU can certainly provide a good orientation value. In the other case with larger data records, PARENT is a more appropriate method to estimate the occurrence of phenomena such as hallucinations, divergences, and omissions. It is thus better suited for the evaluation of DTG models. Nevertheless, the combination of the BLEU, ROUGE, and PARENT scores give a good quantification of the model's capabilities and allow a potentially differentiated conclusion.

3 STUDIES

3.1 Study 1: Conceptual Methodological Replication

The first study aims to conduct a conceptual, methodological replication of Chen et al. (Chen et al., 2019). To get an intuition for how the model works, we replicate one of the training settings presented in (Chen et al., 2019). For evaluation, the authors have only used the methods BLEU and ROUGE. We additionally evaluate with PARENT.

Experimental Setup. Chen et al. (Chen et al., 2019) use WikiBio as data set and they further create two more sets according to the same principle and structure for the domains *Books* and *Songs* by crawling Wikipedia. In total the data sets contain 6452 instances for $Wiki_{books}$, 14787 instances for *WikiBio* and 13079 instances for $Wiki_{songs}$. Table 1 shows the statistical properties of the respective input tables and target summaries.

For optimizing, Chen et al. (Chen et al., 2019) used the *Adam* (Kingma and Ba, 2014) optimizer algorithm with a learning rate of 0.0003. The field gate l_t is being applied and the copy loss weight λ from the Switch-Policy is set to 0.7. Also, the Dual Attention mechanism is applied. The PARENT-λ weight is set to 0.5 for all studies. The model's hyperparameters are set to a hidden size of 500, a field embedding size of 768, and a position embedding of size 5. The number of epochs is not specified explicitly in the paper (Chen et al., 2019); however, the number of epochs in the original code is set to 5000. Due to computational limitations, we have run 330 *epochs* for all studies.

Results. The results of the evaluation of our replication and the values published by Chen et al. (Chen and Mooney, 2008) are quite close. Table 2 shows our values for BLEU-4 and ROUGE-4 and Table 3 those of Chen et al (Chen et al., 2019). Although the results are not identical, we can claim that the values differ within a normal range due to different sampled data records and training time.

Since the PARENT score for the original results for Chen et al. are not available, they cannot be compared. However, it can be observed that in our study, in the evaluations with BLEU-4 and ROUGE-4, the $Wiki_{songs}$ set performs best, and when table information is included in the evaluation process as in PARENT, the $Wiki_{books}$ set performs best. We suspect that the reason for the poorer performance of the *WikiBio* set is that the target summaries may be more diverse and/ or show more divergences which has been shown by Dhingra et al. (Dhingra et al., 2019). This is less likely the case for the $Wiki_{books}$ and $Wiki_{songs}$ sets. While the BLEU-4 and ROUGE-4 scores can only tell us to what degree prediction and target text match, the PARENT scores are more relevant in the context of DTG. Interpreting the PARENT score, we can say that the reference text or the table entails 66.8% of all n-grams from the total predictions. However, the prediction only contains about one-fifth of the information from the table and the target text.

Figure 2 shows the copy loss during the training. While the copy loss for the data sets $Wiki_{books}$ and $Wiki_{songs}$, after a common initial descent, settles at a significantly lower level, the copy loss for the *WikiBio*

Table 1: *Study 1:* Statistics of table properties of Wiki sets.

	Set	Cells			Rows			Columns (Attributes)		
		Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Train	*WikiBio*	5	82	15.3	1	2	1.1	5	80	13.9
	$Wiki_{books}$	3	18	10.6	1	1	1	3	18	10.6
	$Wiki_{songs}$	2	18	9.1	1	1	1	2	18	9.1
Valid	*WikiBio*	2	18	10.5	1	1	1	2	18	10.5
	$Wiki_{books}$	3	52	16.7	1	2	1.2	3	40	13.9
	$Wiki_{songs}$	1	20	9.4	1	1	1	1	20	9.4

Table 2: *Study 1:* Results for the three data set variations obtained after 330 epochs with 200 training instances. Architecture and parameters as in (Chen et al., 2019). Data sets used for training were randomly sampled from original set.

Set	BLEU-4	ROUGE-4	PARENT		
		F-Score	precision	recall	F-Score
$Wiki_{books}$	35.2	21.7	**66.8**	**36.1**	**44.3**
WikiBio	33.4	16.6	61.7	26.1	34.1
$Wiki_{songs}$	**36.9**	**26.8**	66.2	33.5	43.9

Table 3: *Study 1:* Evaluation results for Wiki from Chen et al. (Chen et al., 2019) in 5000 epochs.

Set	BLEU-4	ROUGE-4
		F-Score
$Wiki_{books}$	**37.9**	25.0
WikiBio	36.1	22.1
$Wiki_{songs}$	30.4	**30.1**

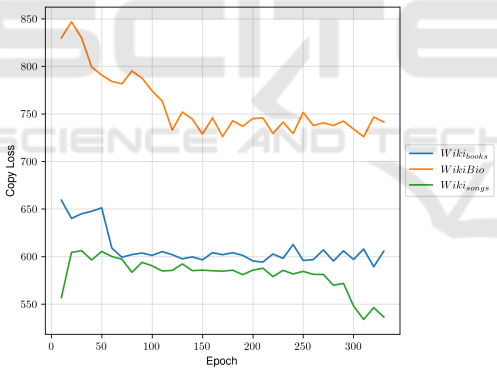

Figure 2: *Study 1:* Development of the copy loss.

remains at a higher value. The copy loss indicates the performance of the copy probability term p_{copy} that aims to learn when to generate over vocabulary and when to copy a word from the table for the prediction. In preprocessing, values from the input table were matched with the target text. During training, the copy probability was maximized at these positions. High or fluctuating values in copy loss thus mean that the model has problems generating a suitable mapping between the behavior of copying values from the table and generating new over vocabulary. The behavior of the copy loss for WikiBio again indicates that it may contain more divergences than the other two sets.

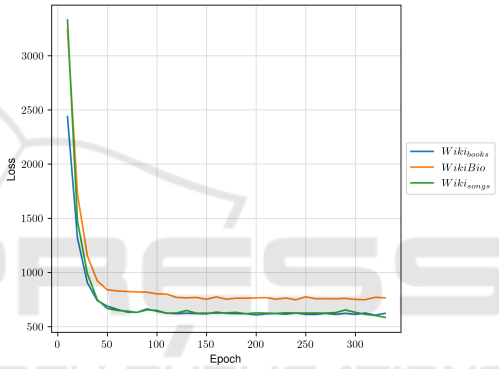

Figure 3: *Study 1:* Development of the loss.

The overall model loss, (c.f. Figure 3), shows a similar picture regarding the behavior of *WikiBio* to $Wiki_{books}$ and $Wiki_{songs}$. While all sets show an initial substantial decrease, *WikiBio* remains at a higher level than the other two sets. This can be observed very well at epoch 250: in Figure 2 for the copy loss and Figure 3 for the loss *WikiBio* shows a slight, simultaneous increase. However, there seem to be other mechanisms that weaken the relatively strong fluctuation of the copy loss.

Discussion. Overall, we can claim that we were able to reproduce the results of Chen et al. (Chen et al., 2019) on sets *WikiBio*, $Wiki_{books}$, and $Wiki_{songs}$. Although the values do not fully match their results, we believe that this may be due to the training duration of only 330 epochs and slight variations between our sampled sets and the authors' sets. Although, according to the BLEU-4 and ROUGE-4 scores, in our study $Wiki_{songs}$ achieves best results, the best performing set according to the PARENT evaluations matches with the best scores of Chen et al. (Chen

et al., 2019) for set $Wiki_{books}$. However, since no PARENT scores are available for the original results of Chen et al. (Chen et al., 2019), we can only speculate whether the best values in the BLEU-4 and ROUGE-4 scores would also be reflected in PARENT.

Concerning our sets, we strongly hypothesize that there are more divergences in *WikiBio* than in the other two. We infer this from the combination of the worse performance in the evaluation scores and the course of the copy loss. The values here are clearly higher and indicate that the model has problems finding a function for the copy or generation behavior of tokens for the prediction. At these points, hallucinations arise because it learns to use supposedly contextless content words. This, in turn, leads to a lower score in the precision because the table or the target summary entails fewer n-grams of the prediction. These developments in the copy loss are reflected in the overall loss of the model. While $Wiki_{books}$ and $Wiki_{songs}$ remain at a lower level, the loss of WikiBio stands out from them.

The results of this study engage the intention to apply the model to a more diverse data set. On the one hand, to test the generalization ability when domains are no longer tested separately from each other, but also to observe the model's behavior with more diverse and larger input tables. In addition, the results suggest that PARENT scores provide a more nuanced view of the performance of a DTG model.

3.2 Study 2: Methodological Replication with ToTTo

The second study aims to apply the model to a more diverse data set. More diverse in terms of the topics covered but also in terms of the type of records. ToTTo contains statistics about, e.g., sports, elections, and the sciences. Such data requires a certain degree of deduction skills and the ability to generalize. Also, unlike Chen et al. (Chen et al., 2019), the different domains are no longer trained separately to examine generalizability.

Experimental Setup. The ToTTo data set offers the possibility to include metadata to different degrees in a model's training. We perform a total of six runs which are divided into two main categories: (1) sets containing only direct table information such as head and subtitle marked by $_{TT}$ and (2) table information as in $_{TT}$ plus page title denoted by $_{PT}$. Since the ToTTo data set comes entirely from Wikipedia entries, page title in this context means the title of the entire Wikipedia entry. For both sets there are three training sets: (1) $few\text{-}shot$ with 200 training instances, (2) *extended* with 400 training instances and (3) *standard* with the set split into Train (10%), Valid (10%) and Test (80%). Tables 4 and 5 show the statistical properties of the respective input tables and target summaries. Compared to the Wiki sets, the model covers with tables 28 times larger and an decreased number of tokens per target summary (20 vs. 18).
All sets of a category ($_{TT}$, $_{PT}$) consist of the same subset of records for a better estimation of the effect of the training size. In total, our studies include 42,452 data records from the ToTTo data set. We excluded records with multiple row and column spans to keep the format of the tables to a minimum. We also excluded tables that contain field values with more than 20 tokens to avoid tables that contain text rather than data points.

We only considered the edited, "final" description in all sets, as we believe that hallucinations can best be avoided with a clean table-description alignment.

For the rest, we use the same hyperparameters as in study 1. Again, we trained all sets for 330 epochs. Dual Attention, Switch-Policy, and field gate are applied, the copy loss weight λ is again set to 0.7, and the PARENT λ weight is set to 0.5. The learning rate remains with a step size of 0.0003 along with the Adam optimizer.

Results. We first analyze the evaluation results of all six sets. Table 6 shows that the two standard variations of the ToTTo set show the best values except for a small deviation in the ROUGE score. We observe that scores increase with the number of training instances. Additionally, they also increase with the metadata added to the set. But only in tendency, as the values for the Few-Shot setting with less metadata ($ToTTo_{few\text{-}shot\_TT}$) are better than those for the setting with more metadata ($ToTTo_{few\text{-}shot\_PT}$).

All in all, there is a discrepancy between the BLEU-4 precision values and the PARENT precision values. It shows that the models recognize that the prediction values must be taken from the table (PARENT precision) but do not select the same content as the target summary (BLEU-4).

The PARENT recall shows that only about 1.2 - 3.0% of the n-grams from the table and target text appear in the predictions at all[2]. So, although on average about 45% of the n-grams of all predictions are entailed by the table and target summary, they, on average, contain only about 1.5% of all n-grams occurring in the table and target text.

The BLEU-4 values for the respective variants $few\text{-}shot$ and *extended* remain below a certain level during the training. The two *standard* variations, on

[2]With the PARENT λ weight set to 0.5.

Table 4: *Study 2:* Statistics of token in tables and target texts of ToTTo sets.

	Set	Token in table			Token in target text		
		Min	Max	Mean	Min	Max	Mean
Train	$ToTTo_{few\text{-}shot\ TT}$	8	7752	426.9	6	52	17.8
	$ToTTo_{few\text{-}shot\ PT}$	9	3678	351.6	4	67	17.3
	$ToTTo_{extended\ TT}$	8	9462	417.0	6	52	17.7
	$ToTTo_{extended\ PT}$	9	3802	329.5	4	67	17.7
	$ToTTo_{standard\ TT}$	5	12623	388.8	4	60	17.6
	$ToTTo_{standard\ PT}$	7	12250	391.5	4	67	17.5
Valid	$ToTTo_{few\text{-}shot\ TT}$	5	9943	392.5	4	59	17.5
	$ToTTo_{few\text{-}shot\ PT}$	8	12250	377.8	4	59	17.5
	$ToTTo_{extended\ TT}$	5	12623	382.9	4	60	17.5
	$ToTTo_{extended\ PT}$	7	12250	394.5	4	61	17.5
	$ToTTo_{standard\ TT}$	5	16920	396.5	4	67	17.3
	$ToTTo_{standard\ PT}$	6	14029	380.6	4	61	17.3

Table 5: *Study 2:* Statistics of table properties of ToTTo sets.

	Set	Cells			Rows			Columns (Attributes)		
		Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Train	$ToTTo_{few\text{-}shot\ TT}$	5	4048	198.3	1	513	25.1	4	17	7.9
	$ToTTo_{few\text{-}shot\ PT}$	4	1938	161.0	1	323	18.5	4	19	8.7
	$ToTTo_{extended\ TT}$	4	5451	202.9	1	908	26.0	4	27	7.8
	$ToTTo_{extended\ PT}$	4	3391	154.9	1	323	17.6	4	19	8.8
	$ToTTo_{standard\ TT}$	2	6085	180.2	1	908	23.4	2	39	7.7
	$ToTTo_{standard\ PT}$	3	8755	179.5	1	1250	20.4	3	39	8.8
Valid	$ToTTo_{few\text{-}shot\ TT}$	2	5597	185.6	1	908	24.1	2	34	7.7
	$ToTTo_{few\text{-}shot\ PT}$	3	8755	177.8	1	1250	20.2	3	35	8.8
	$ToTTo_{extended\ TT}$	2	6085	177.1	1	908	23.0	2	39	7.7
	$ToTTo_{extended\ PT}$	3	8755	181.3	1	1250	20.6	3	39	8.8
	$ToTTo_{standard\ TT}$	2	10352	184.0	1	1529	23.9	2	37	7.7
	$ToTTo_{standard\ PT}$	3	10353	176.6	1	862.8	20.3	3	40	8.7

the other hand, outperform the other sets, and the variation with the most metadata ($ToTTo_{standard\_PT}$) achieves higher values from epoch 90 onwards than the variation with less metadata ($ToTTo_{standard\_TT}$).

The copy loss history of all ToTTo variants is shown in Figure 4. The variants with more training instances tend to achieve better scores (i.e., lower values) than those with fewer training instances. And within these spaces, the sets with more metadata again achieve lower scores than those with less metadata. As can be seen in Figure 4, a large part of the summaries contains information from the page title and therefore $ToTTo_{standard\_PT}$ performs significantly better than $ToTTo_{standard\_TT}$. Since all sets have the same subset of records, this can be stated reliably. Nevertheless, it is interesting to see that $ToTTo_{extended\_PT}$ performs on about the same level as $ToTTo_{standard\_TT}$. Probably this is also related to the additional information of the page title.

Figure 5 shows the overall loss of the models. In general, the values for the loss decrease visibly for all sets. One can see very well that these developments are characteristic for the respective variant pairs ($_{TT}$ and $_{PT}$). The *few-shot* variants show the steepest decline, followed by the *extended* variants and finally the *standard* variations. This is due to the fact that it is more difficult for the model to find a mapping function for in- and output the more instances are used for training.

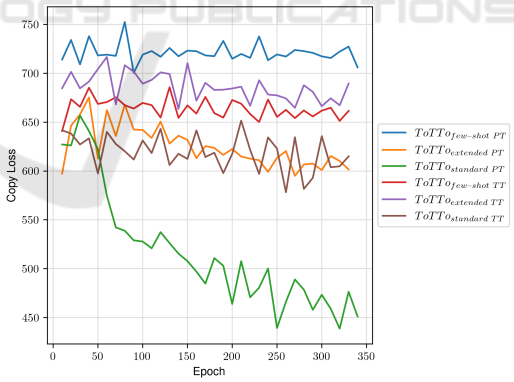

Figure 4: *Study 2:* copy loss with ToTTo. The suffix $_{TT}$ denotes data sets with only additional table information and $_{PT}$ data sets with additional page and table information.

Discussion. In the evaluation results of study 2, we cannot find any contradictory scores between BLEU-

Table 6: *Study 2:* Results for the ToTTo data set variations obtained after 330 epochs with varying amount of training instances. The suffix $_{TT}$ denotes data sets with additional table information and $_{PT}$ those with additional page and table information.

Set	BLEU-4	ROUGE-4	PARENT		
		F-Score	precision	recall	F-Score
$ToTTo_{few\text{-}shot\_TT}$	4.7	1.1	40.5	2.8	4.0
$ToTTo_{extended\_TT}$	5.2	1.3	42.2	3.1	4.5
$ToTTo_{standard\_TT}$	**8.3**	**2.6**	**51.4**	**4.9**	**7**
$ToTTo_{few\text{-}shot\_PT}$	3.4	0.7	38.1	2.4	3.4
$ToTTo_{extended\_PT}$	5.2	**1.2**	43.2	3.1	4.5
$ToTTo_{standard\_PT}$	**9.6**	0.02	**55.4**	**6.1**	**8.7**

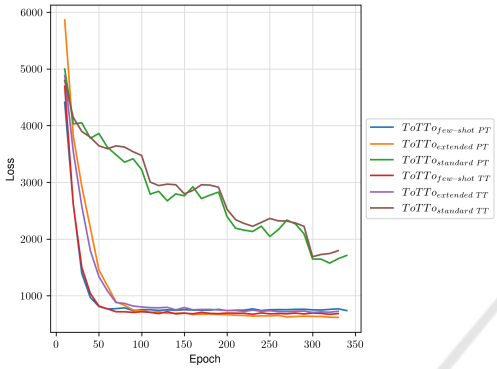

Figure 5: *Study 2:* loss with ToTTo. The suffix $_{TT}$ denotes data sets with only additional table information and $_{PT}$ data sets with additional page and table information.

4 and PARENT. All passages with the highest PARENT scores also show the highest BLEU-4 score and vice versa. However, the discrepancy between the two scores is much higher than in the evaluation scores of study 1. While the best performing set $ToTTo_{standard\_PT}$ shows a PARENT precision score of about 55%, the BLEU-4 is only 9.4. In contrast, the PARENT precision of *WikiBio* from study 1 shows a value only about seven percentage points higher but has a BLEU-4 score of 33.4. This shows more clearly what was already apparent in study 1. Because the BLEU score does not take the input table into account in the evaluation calculation, we get a distorted picture of the results for our purposes. We do not aim to precisely reproduce a table's target text but to obtain an output text that covers the table to a certain extent and represents the correct representation in a semantic context. In relation to classic DTG data sets such as WeatherGov, WebNLG, and WikiBio, this metric may still be appropriate, as the target texts leave less room for variation as compared to ToTTo.

Regarding the coverage criterion, we can state that the model at least learns that it should take content from the table. This is most noticeable in most metadata and training instances sets.

If we only look at the results of the sets used in this study, we see the same movement in all aspects: the more training instances a set has, the better it performs. If the set also includes more metadata, it performs even better. The Few-Shot setting in our study with ToTTo does not achieve comparable results to those of Chen et al. (Chen et al., 2019).

Ultimately, the development in Figure 5 suggests that the two standard variants as whole produce models that are better able to generalize, i.e., respond to previously unseen input. While the *few-shot* and *extended* variants both move relatively quickly to a low level and stay there, the model takes longer to adapt the mapping between in- and output in the standard variant. Again, it would be interesting for further research to see if the loss of the *standard* variants can settle to the level of the other sets with longer training time.

3.3 Study 3: Methodological Replication with VC-SLAM

For our final study, we apply the model to the VC-SLAM data set. As collection of data records from Open Data Portals, it gives us a first impression of the applicability of the model in the context of the goals of the Open Data Charter. Although the data set is small, it reflects a real-world application in terms of quantity (data record sizes) and quality (unedited and unaligned records and descriptions).

Experimental Setup. This study follows the same settings as study 2 in terms of input representations, input parameters to the model, and evaluation. Since this set is much smaller in comparison to WikiBio and ToTTo, we applied only one variation: Train (59.4%), Valid (29.7%), and Test (10.9%). At VC-Slam, we experience the largest input tables so far, seen in Table 7 and Table 8. The model has to cope with even more input per example than in the other two studies, while training with just over a quarter of the instances of the Few-Shot setting.

Results. Table 9 shows the BLEU-4, ROUGE-4 and PARENT evaluation scores. While *VC-SLAM* achieves a BLEU-4 value of 5.1, the model with only 30 training instances achieves a PARENT precision

Table 7: *Study 3:* Statistics of token in target texts of *VC-SLAM*.

	Set	Token in tables			Token in target texts		
		Min	Max	Mean	Min	Max	Mean
Train	*VC-SLAM*	80	1940850	96338.2	10	367	73.4
Valid	*VC-SLAM*	605	982277	59003.4	10	193	56.4

Table 8: *Study 3:* Statistics of table properties of *VC-SLAM*.

	Set	Cells			Rows			Columns (Attributes)		
		Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Train	*VC-SLAM*	13	934590	32711.7	3	54975	2947.6	4	28	11.1
Valid	*VC-SLAM*	151	539603	20873.7	13	19985	1546.2	4	33	13.5

of 26.7. So roughly one-quarter of the n-grams of the prediction is entailed by the table and the target summary, and these cover about 0.1% of the entire content.

In Figure 7 we can observe that the overall model loss reaches a low level quickly and stays there. Since the set is minimal, the model has fewer problems finding a mapping between input and output than is the case, for example, with the other sets.

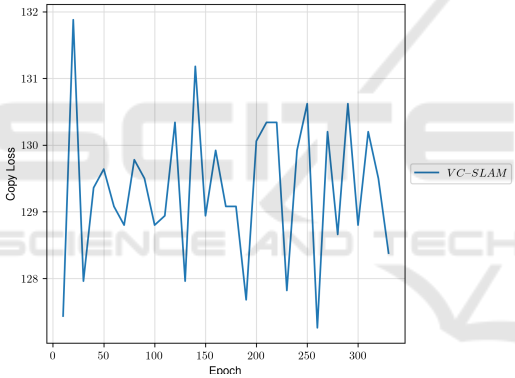

Figure 6: *Study 3:* copy loss with VC-SLAM.

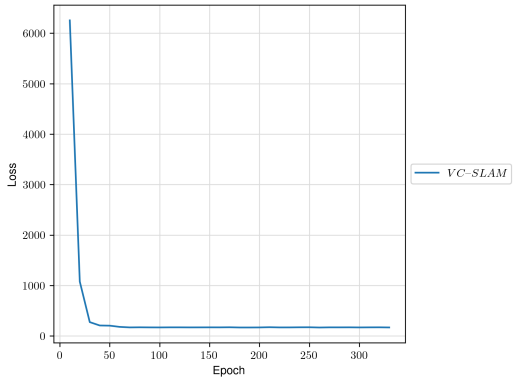

Figure 7: *Study 3:* loss with VC-SLAM.

Discussion. The results of study 3 compare poorly with those of the other studies and are therefore considered in isolation. Even though the model received only 60 training samples, the predictions still achieved a PARENT precision value of almost 27%. This exceeds the expectations we had for this study. Although there are few training instances, the input comes from several domains and has extensive data records. Additionally, the summaries do not have edited properties, so they are not corrected in respect of alignment. The underlying input tables and target texts are larger than those of the other two studies. These findings raise the question of what role the BLEU score and the ROUGE score still play in this application area and whether they are any longer a benchmark for DTG systems. Almost the same applies to the PARENT recall value because its value can only be interpreted to a limited extent in applications with large inputs. For such data sets, a suitable PARENT-λ value should be carefully specified, which defines the relation of content in the summary and that of the table. A mismatch in the PARENT RECALL leads to the F-Score being pulled down by worse recall values.

Overall, the results provide a little perspective on the application in an authentic setting regarding the quality of the data itself rather than quantity.

4 DISCUSSION AND CONCLUSION

This work aimed to evaluate an existing model for DTG that fits our requirements for its applicability in Open Data Portals or the support of Semantic Models for a better organization of the former. The requirements were that it should be domain-independent and not data-hungry. We decided on an architecture that divides the task of generating a coherent text to underlying (semi-) structured data into two independent tasks: content-selection and language generation, with the latter relying on a pre-trained Language Model. In addition, we aimed to find an evaluation

Table 9: *Study 3:* BLEU-4, ROUGE-4 and PARENT scores of VC-Slam after epoch 330.

Set	BLEU-4	ROUGE-4	PARENT		
		F-Score	precision	recall	F-Score
VC-SLAM	5.1	0.2	26.7	0.3	0.5

method that considers both the target text and the associated table in the assessment of the results.

In the first study, the goal was to replicate the results of Chen et al. (Chen et al., 2019). We included the PARENT score for evaluation and were to get additional insights. The *WikiBio* set performed worst in both studies. Using the model outputs and the PARENT score, we can fairly confidently claim that the *WikiBio* data set exhibits more divergences than the other two Wiki sets. Unfortunately, no corresponding data are available for $Wiki_{songs}$ and $Wiki_{books}$, but Dhingra et al. (Dhingra et al., 2019) found that the WikiBio data set in general contains about 62% divergences. The observation that models perform better with data with fewer divergences was confirmed with the help of the ToTTo data set in study 2. Figure 8 shows the development of the copy loss for all sets. Even in comparison with the otherwise better-performing Wiki sets, it can be seen that the $ToTTo_{standard\_PT}$ achieves the best scores. Precise table-description alignment and additional metadata allow the framework to better learn the relationship between copy and generation. Overall, it can be observed that those with more metadata perform better among all settings. In the context of Open Data Portals and the application of DTG for semantic modeling, this inevitably leads to the conclusion that we need metadata to generate metadata for another system. However, it must also be noted that the table-description alignment of ToTTo was done under the premise that title and other direct metadata belong to the actual table (Parikh et al., 2020).

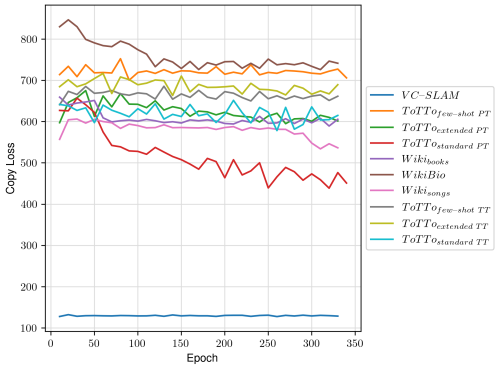

Figure 8: Development of the copy loss with all Data Sets.

Next we look at necessary data sets for DTG. Apart from the Wiki sets, we observed that more training instances perform better. As can be seen in Table 10 in direct comparison, the values improve more or less proportionally to the size of the training set. Despite the pre-trained language model, the task of copying does not seem as trivial as Chen et al. (Chen et al., 2019) assumed. That the ability of content-selection "can be learned by reading a handful of tables" (Chen et al., 2019, p. 1) does not hold in our studies. This may be true in the case of the Wiki sets, but since the tables in most records are only one line, the task is easier to perform. The situation is different for ToTTo and VC-Slam. The records for *ToTTo* are on average 21 lines, for *VC-SLAM* even over 2400. Also, for this reason, a 1:1 translation of the performance results should be done with reservation. Additional data sets such as ToTTo or an extension of the VC-SLAM set are needed to directly compare architectures and their generations.

For example, Nan et al. (Nan et al., 2020) published a large open-domain DTG data set named *DART* (Data Record to Text). The corpus is enriched by annotated tree ontologies converted from underlying tables. The corpus is constructed from different sources like Wikipedia, WikiSQL and incorporated records from WebNLG. Further comparison and research using these data sets are needed to evaluate their contributions further.

Overall, the results of study 1, especially under the BLEU-4 metric, were the best of all three studies conducted. The results are not surprising considering that these sets have, on average, only one data record. It is to be expected that the target text with such data will entail the majority of the table. Thus the generating system will have less room for maneuver in the prediction, taking the table into account.

In general, we have repeatedly encountered the limits of the validity of the BLEU-4 and ROUGE-4 scores in the context of DTG. While still standard, we have found that two questions should be asked in advance when using or interpreting them: (1) Is the goal that the predictions are as close as possible to the target text?, and (2) Is the goal that the prediction covers as much of the tables or target texts as possible? The former need not always be the goal. Originally, both metrics come from Neural Machine Translation, where the ambition was to get a prediction as close as possible to the target. In the context of DTG, this question certainly also depends on the used data. In our understanding, the objective is different with diverse data. Our goal is a description

Table 10: Overview of all evaluation scores for all Data Sets.

Set	Training instances	BLEU-4	ROUGE-4	PARENT		
				precision	recall	F-Score
$ToTTo_{few\text{-}shot\ TT}$	200	4.7	1.1	40.5	2.8	4
$ToTTo_{few\text{-}shot\ PT}$	200	3.4	0.7	38.1	2.4	3.4
$ToTTo_{extended\ TT}$	400	5.2	1.3	42.2	3.1	4.5
$ToTTo_{extended\ PT}$	400	5.2	1.2	43.2	3.1	4.5
$ToTTo_{standard\ TT}$	4245	8.3	2.6	51.4	4.9	7
$ToTTo_{standard\ PT}$	4245	9.6	0.02	55.4	6.1	8.7
$Wiki_{books}$	200	35.2	21.7	**66.8**	**36.1**	**44.3**
WikiBio	200	33.4	16.6	61.7	26.1	34.1
$Wiki_{songs}$	200	**36.9**	**26.8**	66.2	33.5	43.9
VC-SLAM	60	5.1	0.2	26.7	0.3	0.5

that reflects semantic concepts and the context of the given table. Apart from that, both the BLEU and the ROUGE score do not include the input table in the evaluation. Furthermore, Wang (Wang, 2020) argues that "hallucinated facts may unrealistically boost the BLEU score. Thus the possibly misleading evaluation results inhibit systems to demonstrate excellence on this task" (Wang, 2020, p. 312).

Concerning the second inquiry, the PARENT recall score is also to be questioned. Again, this depends on the data sets used as just described. However, for data sets with more extensive records, it is rather unlikely to aim for the highest possible coverage of the entire input in the prediction. Above all, such a task description is not available as a reference in the data sets known to us (for larger records like those found in ToTTo and VC-SLAM). In any case, the PARENT-λ weight should be considered in the evaluation, and, if necessary, experiments should be carried out with different proportions of coverage. Finally, there remains the question of factual accuracy. Even if higher PARENT precision scores are achieved, i.e., a certain number of n-grams of the table and the target text are found in the prediction, this does not necessarily mean that the prediction content is factually correct. Especially in the field of journalism or public relations, it must be ensured that no "fake news" is spread (Portet et al., 2009).

The final development of the models, as observed in Figure 9 in the overall comparison, shows that the frameworks, in general, seem to cope with the task of the DTG. As expected, the sets with fewer instances show a faster framework adaptation. However, there is the assumption that these models generalize less well than the standard variations. Future research could confirm this.

In our model, we have used the Language Model GPT-2 with 177 million parameters. Studies by Brown et al. (Brown et al., 2020) have shown that proficiency in in-context learning increases with the parameters of the pre-trained Language Model. For

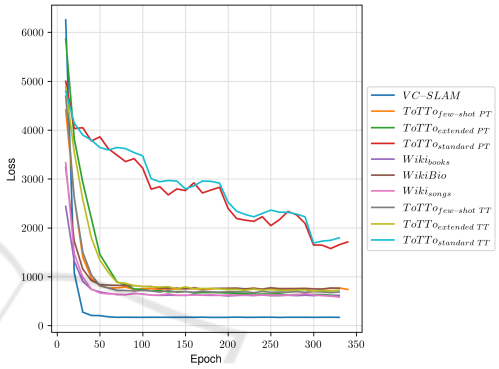

Figure 9: Development of the loss with all Data Sets.

future research, it would be interesting to evaluate this framework with GPT-3, the successor of GPT-2.
It would also be of interest to extend the studies presented here and examine the results after a longer training period. Except for the development in the overall loss of the smaller sets, the values in the BLEU and copy loss indicate that the scores have not yet leveled off and that further development could still occur.

In the course of the last year, other promising DTG systems have been developed. For example, Rebuffel et al. (Rebuffel et al., 2021) follow a word-level approach to control hallucinations in generation. Labels were obtained by employing co-occurrence analysis and dependency parsing. The authors achieved state-of-the-art performances on WikiBio for the PARENT F-Score of approximately 56%. Filippova (Filippova, 2020) reaches on WikiBio up to 52% on the PARENT F-Score by adding a hallucination score as an additional attribute to an instance. However, it can be assumed that the frameworks in other domains or with different data sets than WikiBio will achieve inferior results.

In light of the results from the VC-SLAM data set, we encourage expanding research in this direction and collecting more data from Open Data Portals. Despite the small training set, the predictions al-

ready show a nearly 27% entailment. Although these data records are substantially larger and more diverse, the model seems to adapt. However, concerning our research question, we cannot claim to have obtained satisfactory results with a minimal amount of real-world data. At this point, another research objective emerges, which has already been articulated by Burgdorf et al. (Burgdorf et al., 2020). To say reliably whether given metadata is useful for semantically modeling tabular data requires some kind of assessment or evaluation. The authors propose to use historical data from the (potentially) established ontology to make some kind of prediction about how much manual effort a given semantic model will need with a given data set. Schauppenlehner and Muhar (Schauppenlehner and Muhar, 2018) support this approach.

Finally, our conclusion is somewhat ambivalent. We were able to test the present framework on an open-domain setting and achieved valuable results, even if not in the few-shot setting. However, many open research questions remain: How can the generations of DTG systems be qualitatively evaluated? We have been able to identify a method that allows us to assess the degree of entailment, but this does not tell us anything about factual correctness, nor whether it is semantically relevant for use in the context of Open Data Portals. Furthermore, our results show that the amount of metadata is crucial for the performance of a DTG model. If not all the information needed for the generation can be obtained from the table, we must rely on additional information. At this point, a vicious circle arises because, in order to generate metadata, we need metadata.

REFERENCES

(2013). G8 open data charter and technical annex.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Burgdorf, A., Paulus, A., Pomp, A., and Meisen, T. (2022). Vc-slam—a handcrafted data corpus for the construction of semantic models. *Data*, 7(2):17.

Burgdorf, A., Pomp, A., and Meisen, T. (2020). Towards nlp-supported semantic data management. *arXiv preprint arXiv:2005.06916*.

Chandola, T. and Booker, C. (2022). *Archival and Secondary Data*. SAGE.

Chen, D. L. and Mooney, R. J. (2008). Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135.

Chen, Z., Eavani, H., Chen, W., Liu, Y., and Wang, W. Y. (2019). Few-shot nlg with pre-trained language model. *arXiv preprint arXiv:1904.09521*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhingra, B., Faruqui, M., Parikh, A., Chang, M.-W., Das, D., and Cohen, W. W. (2019). Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Filippova, K. (2020). Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lebret, R., Grangier, D., and Auli, M. (2016). Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nan, L., Radev, D., Zhang, R., Rau, A., Sivaprasad, A., Hsieh, C., Tang, X., Vyas, A., Verma, N., Krishna, P., et al. (2020). Dart: Open-domain structured data record to text generation. *arXiv preprint arXiv:2007.02871*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Parikh, A. P., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., and Das, D. (2020). Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., and Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rebuffel, C., Roberti, M., Soulier, L., Scoutheeten, G., Cancelliere, R., and Gallinari, P. (2021). Controlling hallucinations at word level in data-to-text generation. *arXiv preprint arXiv:2102.02810*.

Schauppenlehner, T. and Muhar, A. (2018). Theoretical availability versus practical accessibility: The critical role of metadata management in open data portals. *Sustainability*, 10(2):545.

Tygel, A., Auer, S., Debattista, J., Orlandi, F., and Campos, M. L. M. (2016). Towards cleaning-up open data portals: A metadata reconciliation approach. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, pages 71–78. IEEE.

Wang, H. (2020). Revisiting challenges in data-to-text generation with fact grounding. *arXiv preprint arXiv:2001.03830*.