NUIG-DSI's submission to The 💎 GEM Benchmark 2021

Nivranshu Pasricha[1], **Mihael Arcan**[2] and **Paul Buitelaar**[1,2]
[1]SFI Centre for Research Training in Artificial Intelligence
[2] Insight SFI Research Centre for Data Analytics
Data Science Institute, National University of Ireland Galway
n.pasricha1@nuigalway.ie

Abstract

This paper describes the submission by NUIG-DSI to the GEM benchmark 2021. We participate in the modeling shared task where we submit outputs on four datasets for data-to-text generation, namely, DART, WebNLG (en), E2E and CommonGen. We follow an approach similar to the one described in the GEM benchmark paper where we use the pre-trained T5-base model for our submission. We train this model on additional monolingual data where we experiment with different masking strategies specifically focused on masking entities, predicates and concepts as well as a random masking strategy for pre-training. In our results we find that random masking performs the best in terms of automatic evaluation metrics, though the results are not statistically significantly different compared to other masking strategies.

1 Introduction

The GEM Benchmark (Gehrmann et al., 2021) is a living benchmark focusing on generation, evaluation and metrics for a variety of natural language generation tasks including summarization, simplification, dialog and data-to-text generation. In general, the field of natural language generation (NLG) is concerned with automatic generation of human understandable texts, typically from a non-linguistic or textual representation of information as input (Reiter and Dale, 2000). Traditionally, most applications for NLG have relied on rule-based systems designed using a modular pipeline approach (Gatt and Krahmer, 2018). However, recently approaches based on neutral networks with an encoder-decoder architecture trained in an end-to-end fashion have gained popularity. These typically follow the paradigm of pre-training on a large corpus followed by fine-tuning on a task specific dataset and have been shown to achieve state-of-the-art results on many natural language tasks (Raffel et al., 2020; Lewis et al., 2020). When evaluated by human annotators, neural models for data-to-text generation have been found to produce fluent text though such models might struggle in terms of data coverage, relevance and correctness where rule-based systems score high (Castro Ferreira et al., 2020).

In our participation in the GEM benchmark, we submit outputs for four datasets including DART (Nan et al., 2021), WebNLG (Gardent et al., 2017; Castro Ferreira et al., 2020), E2E (Novikova et al., 2017; Dušek et al., 2019) and CommonGen (Lin et al., 2020). We use the pre-trained T5-base model architecture (Raffel et al., 2020) for our submission implemented using the transformers library from Hugging Face (Wolf et al., 2020). We first train on monolingual data before fine-tuning on the task-specific dataset. For DART and WebNLG, we use abstracts from DBpedia (Auer et al., 2007) for training while for the other two datasets, we use monolingual target-side references for pre-training with a masked language modeling objective. We experiment with different masking strategies where we mask entities and predicates (for DART), meaning representation fields (for E2E) and concepts (for CommonGen) and compare the results with commonly used approach of random masking. Our results suggest that random masking achieves the best scores for automatic evaluation metrics for DART, WebNLG and E2E while additional pre-training appears to hurt the performance for CommonGen.

2 Methodology

In this section we define our methodology on the four datasets where we make a submission and subsequently discuss the results based on automatic evaluation metrics defined in the GEM benchmark.

Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 148–154

Tripleset	Antioquia Department	country	Colombia
	Bandeja paisa	ingredient	Chorizo
	Bandeja paisa	region	Antioquia Department
linearisation	Antioquia Department country Colombia Bandeja paisa ingredient Chorizo Bandeja paisa region Antioquia Department		
tags	<*SUB*> Antioquia Department <*PRED*> country <*OBJ*> Colombia <*SUB*> Bandeja paisa <*PRED*> ingredient <*OBJ*> Chorizo <*SUB*> Bandeja paisa <*PRED*> region <*OBJ*> Antioquia Department		
entity types	<*LOCATION*> Antioquia Department <*PRED*> country <*LOCATION*> Colombia <*FOOD*> Bandeja paisa <*PRED*> ingredient <*SAUSAGE*> Chorizo <*FOOD*> Bandeja paisa <*PRED*> region <*LOCATION*> Antioquia Department		
NER tags	<*ORG*> Antioquia Department <*PRED*> country <*GPE*> Colombia <*PERSON*> Bandeja paisa <*PRED*> ingredient <*UNKNOWN*> Chorizo <*PERSON*> Bandeja paisa <*PRED*> region <*ORG*> Antioquia Department		

(a) Additional tags added to the linearised tripleset.

Lexicalisation	Chorizo is an ingredient in Bandeja paisa, a dish from the Antioquia Department region, in Colombia.
Random Masking	Chorizo is an ingredient in Bandeja paisa, a dish **[MASK]** Antioquia Department **[MASK]**, in Colombia.
Entity Masking	**[MASK]** is an ingredient in **[MASK]**, a dish from the **[MASK]** region, in **[MASK]**.
Predicate Masking	Chorizo is an **[MASK]** in Bandeja paisa, a dish from the Antioquia Department **[MASK]**, in Colombia.

(b) Masking strategies for pre-training on monolingual data.

Figure 1: Example of a tripleset from the DART dataset with additional information tags included after linearisation for fine-tuning (top) and different masking strategies applied to a sentence for pre-training (bottom).

2.1 DART

DART (Nan et al., 2021) consists of open domain data records structured in the form of triples paired with crowd-sourced textual annotations in English describing those triples. The data is collected from multiple different sources including tables from Wikipedia, questions from WikiSQL and merged with two existing data-to-text datasets, namely, WebNLG (en) (Gardent et al., 2017) and cleaned E2E (Dušek et al., 2019).

Since both DART and WebNLG are concerned with the task of triple-to-text generation and have the same input data structure, we follow the same approach as defined in Pasricha et al. (2020) for the WebNLG+ challenge. We use the pre-trained T5 model architecture and first train it on a corpus of abstracts from DBpedia with a masked language modeling objective. For masking, we adopt the commonly used approach of randomly masking 15% of the tokens in texts. We further compare this with an approach where we specifically mask only the entities or only the predicates or a combination of both as shown in Figure 1(b). The abstracts are downloaded from DBpedia for the entities which are present in the triples contained in the training set of the DART dataset. Since we did not find an abstract for each unique entity in the training set, we ended up with 9,218 abstracts consisting on 1,654,239 tokens and 83,583 types in total with an average of 179.45 tokens per abstract. After pre-training, we fine-tune on the DART training set to predict the target text conditioned on the linearised tripleset.

	BLEU	METEOR	ROUGE-L
baseline	46.10	37.24	59.61
masked pre-training			
random masking	47.16	37.51	59.99
entity masking	45.92	37.14	59.56
predicate masking	46.73	37.36	59.79
entity + predicate	46.37	37.23	59.51

Table 1: Results from automatic evaluation on the DART validation set with different masking strategies on DBpedia abstracts for pre-training using the T5-small model.

For fine-tuning we linearise the input tripleset into a sequence without modifying the order of the triples in the input. We incorporate additional information to mark the subject, predicate and object in each triple in the input by using <*SUB*>, <*PRED*> and <*OBJ*> tags respectively. Additionally, we also include tags for the type of an entity using DBpedia as shown in Figure 1(a). In the instances where we do not find the type of an entity on DBpedia, we

check whether it refers to a time period or a date and assign it the type <*TIMEPERIOD*>. Otherwise, we assign the type <*MEASUREMENT*> to an entity containing a numeric value followed by some text. The type <*NUMERIC*> is assigned to entities which only consist of numeric values and <*UNKNOWN*> to everything else. Furthermore, as a comparison, we add tags for entities using the named entity recognition pipeline from the spaCy library[1]. All of these tags are included as additional special tokens to the model vocabulary.

For our experiments with masking during pre-training on DBpedia abstracts, we use the small variant of the T5 model architecture. This model has approximately 60 million parameters and is much faster to train compared to other larger variants. We use the pre-trained model implementation from Hugging Face's transformers library (Wolf et al., 2020) which consists of 6 layers each in the encoder and decoder with a multi-head attention sub-layer consisting of 8 attention heads. The word embeddings have a dimension of 512 and the fully-connected feed-forward sublayers are 2048-dimensional. Pre-training on DBpedia abstracts is done on a single Nvidia GeForce GTX 1080 Ti GPU for 10 epochs with a batch size of 8 using the Adam optimizer with a learning rate of 0.001. All the other hyperparameter values are set to their default values. Table 1 shows scores for the output generations on the validation set for BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE-L (Lin, 2004). We find random masking to perform the best in terms of automatic evaluation metrics compared to specifically masking entities or predicates, though the results are not statistically significantly different.

Furthermore, in our experiments we compare the results when additional tags are added to the input either as entity types from DBpedia or NER tags from spaCy or just the <*SUB*>, <*PRED*> and <*OBJ*> tags. For this, we use the T5-base model with approximately 220 million parameters. This model consists of 12 layers each in the encoder and decoder with 12 attention heads in each multi-head attention sublayer. The word embeddings are 768-dimensional for this model and feed-forward sublayer is 3072-dimensional. This model is first pre-trained on DBpedia abstracts with a masked language modeling objective where 15% of the tokens are corrupted randomly. For fine-tuning, we train on the DART training set for 10 epochs on a single Nvidia GeForce GTX 1080 Ti GPU with a batch size of 16 and select the checkpoint with the highest BLEU score on the validation set. We set the maximum output sequence length to 50 words and apply beam search during inference with a beam of size equal to 5. Here we find that adding the three <*SUB*>, <*PRED*> and <*OBJ*> tags achieves the best results compared to tags from DBpedia or spaCy though the differences in the automatic evaluation results are again not statistically significant. For our final submission to the GEM benchmark, we submit the outputs from this model which is fine-tuned with the added <*SUB*>, <*PRED*> and <*OBJ*> tags.

[1] https://spacy.io

	BLEU	METEOR	ROUGE-L
baseline	51.06	40.23	60.86
tags	51.71	40.68	61.10
DBpedia types	50.75	40.33	60.45
spaCy NER	51.05	40.42	61.30

Table 2: Results from automatic evaluation on the DART validation set with different tags for fine-tuning. The results are shown here using the T5-base model which is first pre-trained with the random masking on a corpus of DBpedia abstracts.

2.2 WebNLG

WebNLG (Gardent et al., 2017) introduced the task of RDF-to-Text generation focused on generating a verbalisation in a human language in the output based on a set of RDF-triples in the input. The WebNLG corpus consists of data units made up of RDF-triples extracted from DBpedia (Auer et al., 2007) and paired with reference text lexicalisations. These texts were collected using crowd-sourcing and contain sequences of one or more short sentences in English, verbalising the data units in the input. The first version of the corpus contained triplesets from 15 DBpedia categories and is divided into two subsets, *seen* and *unseen* for evaluation. The ten *seen* categories are *Airport*, *Astronaut*, *Building*, *City*, *ComicsCharacter*, *Food*, *Monument*, *SportsTeam*, *University* and *WrittenWork* and the five *unseen* categories are *Artist*, *Athlete*, *CelestialBody*, *Company*, *MeanOfTransportation* and *Politician*. WebNLG+ (Castro Ferreira et al., 2020) was further introduced to include Russian as another output language and added the category *Company* to the training set as well as three categories *Film*, *MusicalWork* and *Scientist* to the test set.

	BLEU	METEOR	ROUGE-L
baseline	33.73	36.52	53.72
masked pre-training			
MR masking	34.09	36.62	53.64
random masking	34.21	36.50	53.85

Table 3: Results from automatic evaluation on the E2E validation set with different masking strategies on monolingual data for pre-training using the T5-base model.

	BLEU	METEOR	ROUGE-L
baseline	28.94	31.03	55.78
masked pre-training			
concept masking	27.81	29.61	54.87
random masking	26.87	29.83	54.17

Table 4: Results from automatic evaluation on the CommonGen validation set with different masking strategies on monolingual data for pre-training using the T5-base model.

Since the entire WebNLG (en) corpus is already included the DART dataset without any modifications, we use the same model as defined in §2.1 without any further fine-tuning to generate outputs on the WebNLG (en) dataset. Our overall approach is same as Pasricha et al. (2020) for the WebNLG+ challenge 2020 except here we use additional 6,678 DBpedia abstracts for pre-training and the larger DART dataset for fine-tuning which results in a higher scores for automatic evaluation metrics.

2.3 E2E

E2E (Novikova et al., 2017) is concerned with generating texts for a dialogue system from meaning representations (MR) in the restaurant domain. It was introduced with the aim of motivating research in domain-specific end-to-end data-driven natural language generation systems. The input for E2E comprises of meaning representations with up to 8 different fields including *name*, *near*, *area*, *food*, *eatType*, *priceRange*, *rating* and *familyFriendly* while the output comprises of sentences typically made of up 20 – 30 words in English verbalising the input.

We follow the same approach as described in §2.1 and experiment with masking strategies for pre-training on monolingual data. Instead of using additional out-of-domain data, we use the target side references from the E2E dataset for pre-training with a masked language modeling objective. Here we compare the results on two masking strategies, one where we mask 15% of the token spans randomly and another where we mask specific values based on meaning representation fields such as restaurant names, area, price, etc. This approach is similar to the one described in §2.1 where we masked specifically masked entities and predicates. Table 3 shows scores for the output generations on the validation set for BLEU, METEOR and ROUGE-L. We again find that random masking appears to perform better though the differences in terms of automatic evaluation metrics are not significantly different.

For our submission to the GEM benchmark, we use the same model architecture and hyperparameter values as described previously for DART to generate the output submissions on the E2E test set and challenge sets. This model is first pre-trained on the monolingual target side with a masked language objective where the spans of text are masked randomly and the fine-tuned on the E2E training set containing pairs of meaning representations and target texts.

2.4 CommonGen

CommonGen (Lin et al., 2020) was introduced with the goal of testing state-of-the-art text generation systems for the ability of commonsense reasoning. The task for CommonGen is to generate a coherent sentence in English describing an everyday scenario using a set of concepts such as *man*, *woman*, *dog*, *throw* and *catch*. Lin et al. (2020) have shown that large pre-trained language models are prone to hallucinations and can generate incoherent sentences such as "*hands washing soap on the sink*" for the concept set {*hand*, *sink*, *wash*, *soap*}. Two key challenges identified by the creators of this dataset are *relational reasoning* with underlying commonsense knowledge for given concepts and *compositional generalization* for unseen combinations of concepts.

We again start with the T5-base model and experiment with masked pre-training on the monolingual target side of CommonGen. As described in §2.3 we compare two strategies of masking where we mask spans of text randomly or specifically mask tokens which correspond to concepts in the training set. Table 4 shows scores for the output generations on the validation set for BLEU, METEOR and ROUGE-L. For fine-tuning we shuffle the concepts

Dataset	subset	Metrics (Lexical Similarity and Semantic Equivalence)						
		METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	BLEURT
CommonGen	val	0.310	64.37	33.08	55.78	28.77	0.893	-0.380
	sample	0.304	63.72	32.52	54.82	28.24	0.890	-0.391
DART	val	0.396	72.44	48.75	58.77	49.42	0.916	0.192
E2E clean	val	0.366	72.12	45.70	53.87	34.21	0.909	0.228
	test	0.354	73.23	45.71	53.45	31.74	0.913	0.205
	sample	0.365	71.72	45.39	53.81	34.20	0.910	0.221
	scramble	0.349	72.06	44.32	51.69	30.52	0.910	0.176
WebNLG (en)	val	0.391	76.08	53.59	62.51	52.10	0.931	0.282
	test	0.341	71.41	46.66	57.13	41.43	0.910	0.138
	sample	0.389	75.48	53.00	62.38	51.35	0.929	0.260
	scramble	0.343	71.54	47.02	57.07	41.74	0.909	0.140
	numbers	0.338	70.36	45.98	56.78	41.33	0.909	0.101

Dataset	subset	Metrics (Diversity and System Characterization)								
		MSTTR	Distinct$_1$	Distinct$_2$	H_1	H_2	Unique$_1$	Unique$_2$	$\|\mathcal{V}\|$	Output Len.
CommonGen	val	0.54	0.11	0.37	6.9	10.3	532	2.4k	1.2k	10.9
	sample	0.55	0.16	0.46	6.8	10.0	455	1.6k	862	11.0
DART	val	0.42	0.05	0.15	7.4	9.9	1.3k	5.0k	3.1k	22.7
E2E clean	val	0.26	0.001	0.004	5.6	7.0	11	68	144	23.4
	test	0.27	0.001	0.005	5.7	7.1	5	33	136	22.4
	sample	0.44	0.01	0.027	5.6	7.0	6	43	117	23.7
	scramble	0.47	0.01	0.034	5.7	7.1	7	56	117	22.4
WebNLG (en)	val	0.54	0.10	0.30	8.5	11.9	1.1k	4.8k	3.2k	19.2
	test	0.65	0.04	0.16	8.0	10.9	368	2.1k	1.5k	19.5
	sample	0.57	0.20	0.50	8.3	11.3	942	3.0k	1.9k	19.2
	scramble	0.50	0.11	0.32	7.9	10.6	362	1.5k	2.9k	19.8
	numbers	0.65	0.12	0.32	7.9	10.6	426	1.6k	1.1k	19.6

Table 5: Results from automatic evaluation metrics measuring lexical similarity, semantic equivalence, diversity and system characteristics on the validation set, test set and the three challenge sets – sample, scramble and numbers for DART, WebNLG (en), E2E and CommonGen.

in the input before concatenating them into a single sequence. We find in our results that additional pre-training on monolingual data on the target appears to hurt the performance when measured with automatic evaluation metrics. This is true in both the cases when masking is done randomly or when only specific concepts are masked.

3 Results

Table 5 shows results on the validation set, test set and the challenge sets evaluated using GEM metrics[2]. At the time of writing we do not have access to all the references in the test set as well as the challenge sets for DART and CommonGen, hence scores on some subsets are not shown.

The evaluation metrics are divided into different categories measuring lexical similarity, semantic equivalence, diversity and system characteristics. Popular metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE-1/2/L (Lin, 2004) are used for lexical similarity, while recently proposed metrics such as BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020) which rely on sentence embeddings from pre-trained contextualised embedding models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are used for evaluating semantic equivalence. To account for the diverse outputs, Shannon Entropy (Shannon et al., 1950) is calculated over unigrams and bigrams (H_1, H_2) along with the mean segmented type token ratio over segment lengths of 100 (MSTTR) (Johnson, 1944). Furthermore, the ratio of distinct n-grams over the total number of n-grams (Distinct$_{1,2}$), and the count of n-grams that appear once across the entire test output (Unique$_{1,2}$) is calculated (Li et al., 2018). The size of the output vocabulary ($|\mathcal{V}|$) and the mean length of the generated output texts are reported as system characteristics (Sun et al., 2019).

Compared to the baselines described in the GEM benchmark (Gehrmann et al., 2021), we observe higher scores in our submissions for automatic metrics on the CommonGen and DART datasets while scoring lower on the cleaned E2E and WebNLG (en) datasets especially on the test and challenge subsets for both E2E and WebNLG.

[2] https://github.com/GEM-benchmark/GEM-metrics

4 Conclusion

We presented a description of the system submitted by NUIG-DSI to the GEM benchmark 2021. We participated in the modeling shared task and submitted outputs on four datasets for data-to-text generation including DART, WebNLG (en), E2E and CommonGen using the T5-base model. We first trained this model with monolingual data from DBpedia abstracts and target side references before fine-tuning on respective training datasets. Additionally we experimented with various masking strategies focusing specifically on masking entities, predicates and concepts as well as a random masking strategy for training. We found random masking to perform the best and submit our final outputs using this approach.

Acknowledgments

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223 and co-supported by Science Foundation Ireland under grant number SFI/12/RC/2289 2 (Insight), co-funded by the European Regional Development Fund.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6116–6124.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Murori Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. Dart: Open-domain structured data record to text generation. *arXiv preprint arXiv:2007.02871*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. 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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Nivranshu Pasricha, Mihael Arcan, and Paul Buitelaar. 2020. NUIG-DSI at the WebNLG+ challenge: Leveraging transfer learning for RDF-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 137–143, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*, 1 edition. Cambridge University Press.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Claude E Shannon, Warren Weaver, and Norbert Wiener. 1950. The mathematical theory of communication. *Physics Today*, 3(9):31.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.