The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics

Sebastian Gehrmann,[9] *** Tosin Adewumi,**[20 21] **Karmanya Aggarwal,**[14]
Pawan Sasanka Ammanamanchi,[15] **Aremu Anuoluwapo,**[21 38] **Antoine Bosselut,**[28]

Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 96–120

Dataset	Communicative Goal	Language(s)	Size	Input Type
CommonGEN (Lin et al., 2020)	Produce a likely sentence which mentions all of the source c7ncepts.	en	67k	Concept Set
Czech Restaurant (				

data through a human-in-the-loop approach.

Increasing multilingualism of NLG research. Another potentially harmful choice by benchmark creators is the choice of the languages of the included datasets. It is often assumed that work on

twenty years and the evaluation methodologies dif-

shown in Appendix C.

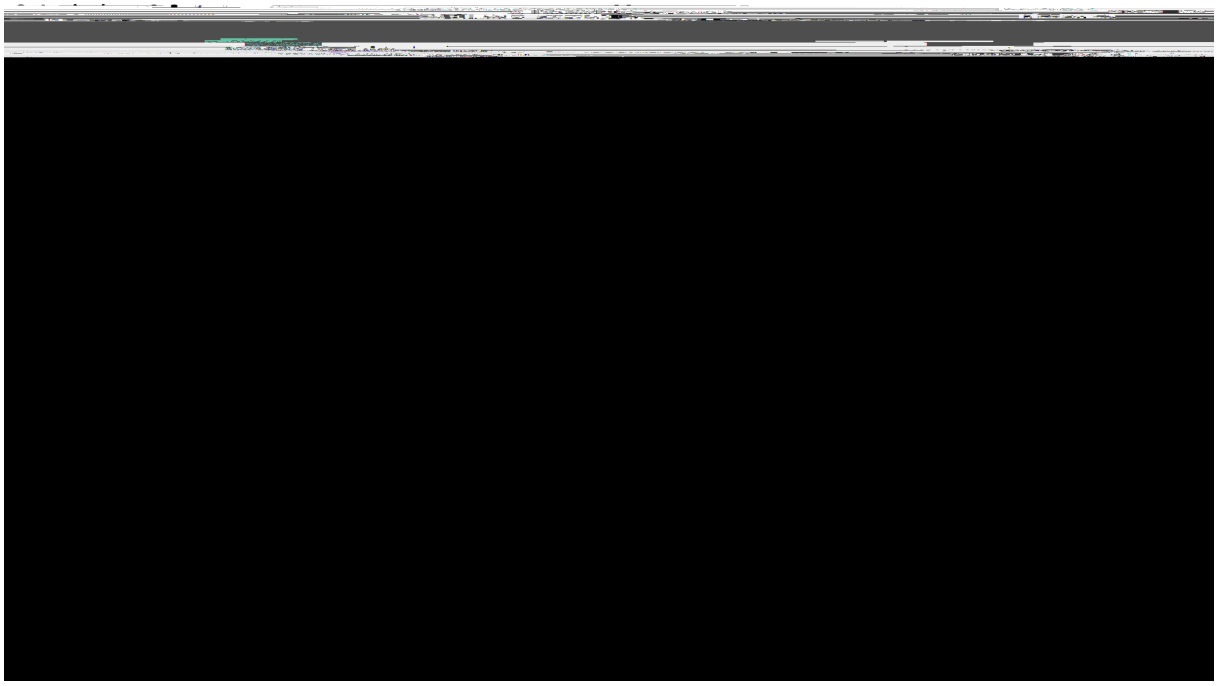
The survey received 28 responses, revealing that

Challenge Set Type	Example	Tasks
Numerical Variation	53 ->79	WebNLG
Attribute Order	English Cheap ->Cheap English	All data-to-text tasks
Typographical Errors	English Cheap ->Enlish Chesp	Schema-Guided, WikiAuto, XSum
No Punctuation	... the dog. ->... the dog	Schema-Guided, WikiAuto, XSum
Backtranslation	fantastic ->toll ->great	Schema-Guided, WikiAuto, XSum
Train & Validation Samples		All tasks
Gender, Ethnicity, Nationality		ToTTo
Input Shape		WebNLG
Syntactic Complexity		WikiAuto
Covid Summaries		MLSUM (es+de), XSum

Table 2: An overview of the types of challenge sets for GEM. The first category are modifications to inputs of a

format of the current cardinal value (e.g. alpha, integer, or floating-point) and replaces the existing value with a new random value as a means to

Dataset	Model	Metrics (Lexical Similarity and Semantic Equivalence)						
		METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	BLEURT
CommonGen	BART	0.301	63.5	32.5	55.1	27.5	0.943	-0.400
	T5	0.291	64.0	29.4	54.5	26.4	0.942	-0.412
Czech Restaurant								



System Foo leads to consistent performance increases in Bar-type metrics on challenges that measure Baz while maintaining equal performance on most met-

All shared task participants will be asked to provide gold annotations on system outputs, which we will then use to evaluate the consistency of crowdsourced annotations.[13]

7 Next Steps

This section lists the currently active developments

interactive exploration of results.

Model Infrastructure. Yacine Jernite wrote the initial script template for evaluating and fine-tuning Hugging Face models with the CommonGen exam-

Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the 13th European*

national Conference on Natural Language Generation, pages 421–426, Tokyo, Japan. Association for Computational Linguistics.

Ondrej Dušek and Filip Jurcıcek. 2016. A context-aware natural language generation dataset for dialogue systems. In *RE-WOCHAT: Workshop on*

Claire Gardent, Anastasia Shimorina, Shashi Narayan,

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kath-

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 and Warren Weaver. 1963. A mathematical theory of communication.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In

22.