trivago

NLG and Evaluation: An overview of current challenges

Saad Mahamood, 25.04.2022

Agenda

1. About Me
2. Overview: What are the issues?
3. Neural NLG Evaluation Efforts
   3.1. GEM Project
   3.2. Generating Challenge Sets
4. Human Evaluation Efforts
5. NLG & Reproducibility Efforts
6. Conclusions

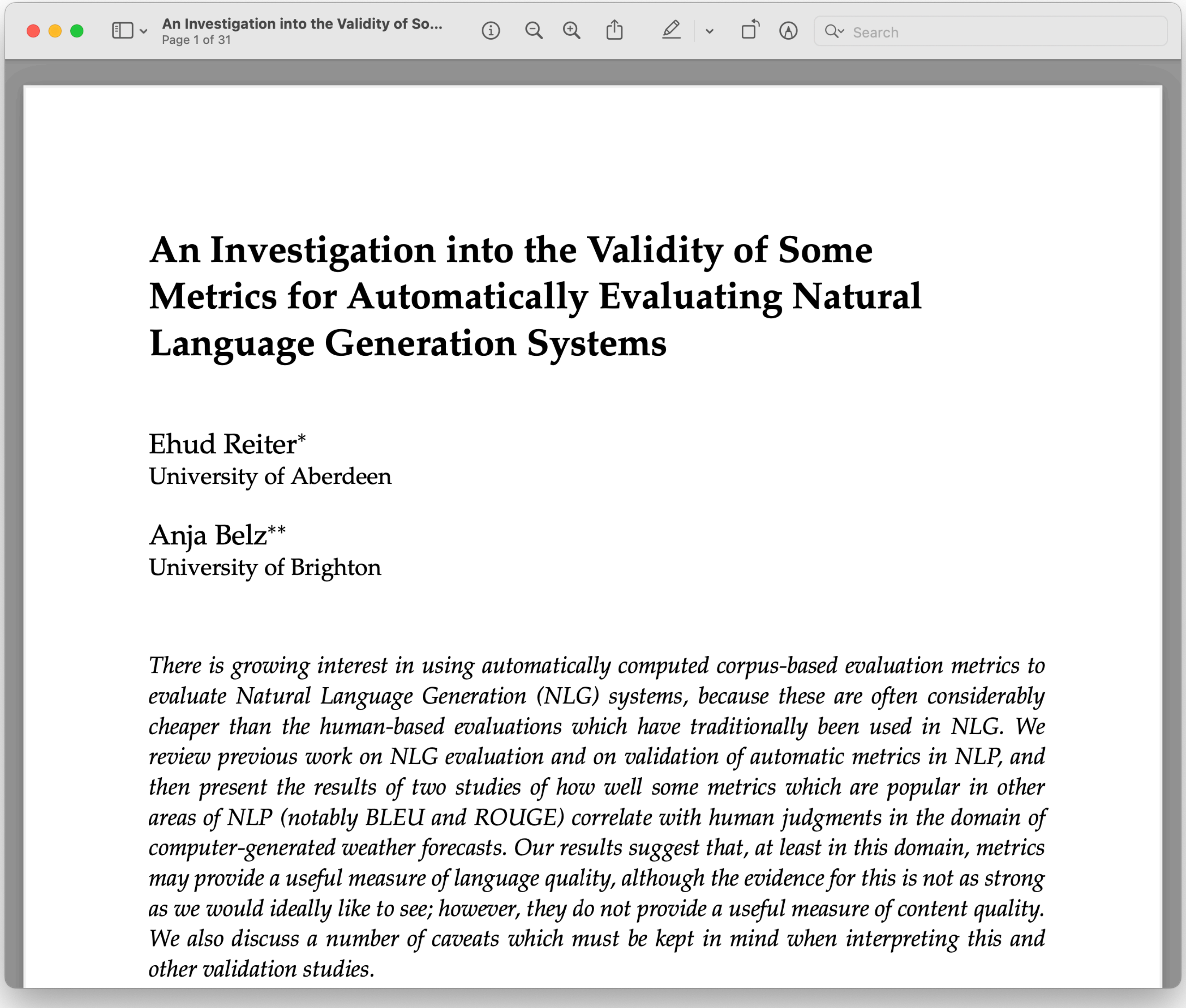
About Me

About Me

- Been active the field of NLG for ~14 years.
- 6 years at Aberdeen University.
    - 4 years PhD
    - 2 years PostDoc
- 5 years at Arria NLG.
- 3 years now at trivago.
    - Released ‘Hotel Scribe’ to generate automated descriptions of accommodations [Mahamood and Zembrzuski, 2019].
    - Lead a team of four data scientists and analysts.
    - Currently, working on more image tagging and data science problems.
    - Actively, participating in NLG research, including research projects. In particular, focused on the topic of evaluations.

My Journey into NLG Evaluation

- ~2010: First aware of issues in NLG evaluation, due to Reiter’s & Belz’s paper on investigating the validity of evaluations [Reiter & Belz, 2009]

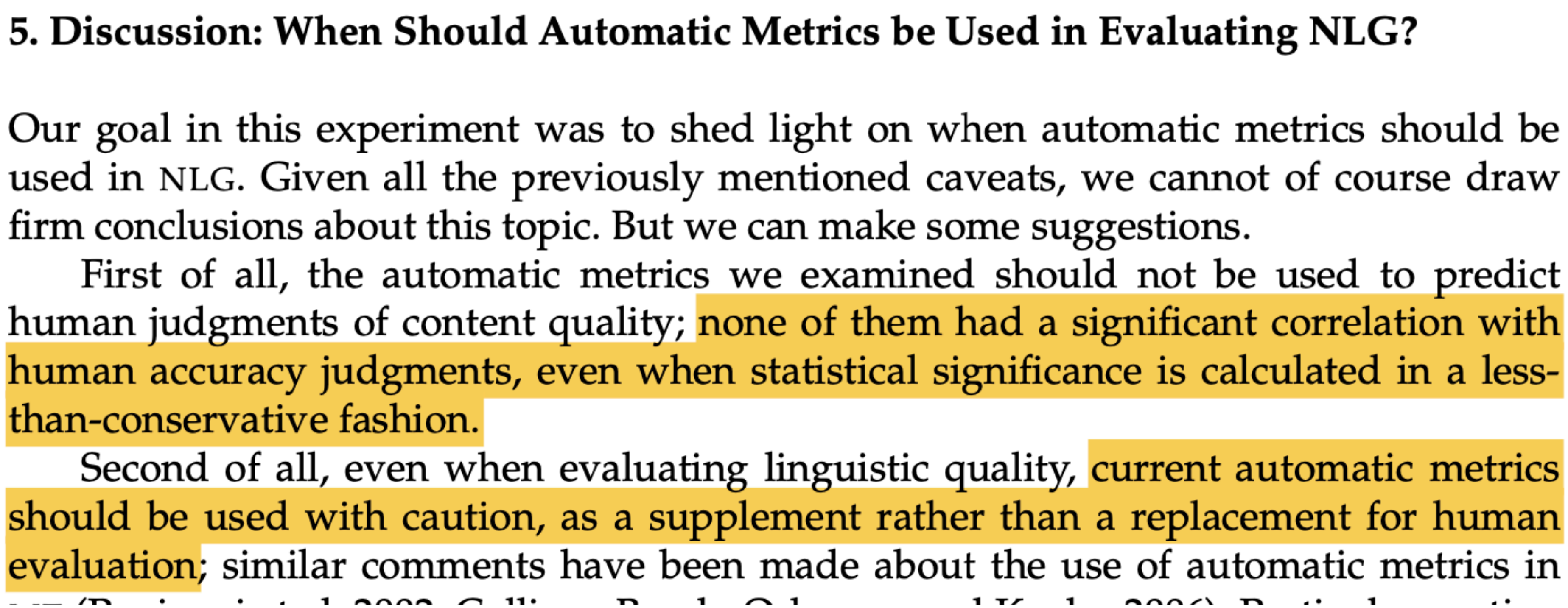


An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems

Ehud Reiter*
University of Aberdeen

Anja Belz**
University of Brighton

There is growing interest in using automatically computed corpus-based evaluation metrics to evaluate Natural Language Generation (NLG) systems, because these are often considerably cheaper than the human-based evaluations which have traditionally been used in NLG. We review previous work on NLG evaluation and on validation of automatic metrics in NLP, and then present the results of two studies of how well some metrics which are popular in other areas of NLP (notably BLEU and ROUGE) correlate with human judgments in the domain of computer-generated weather forecasts. Our results suggest that, at least in this domain, metrics may provide a useful measure of language quality, although the evidence for this is not as strong as we would ideally like to see; however, they do not provide a useful measure of content quality. We also discuss a number of caveats which must be kept in mind when interpreting this and other validation studies.



5. Discussion: When Should Automatic Metrics be Used in Evaluating NLG?

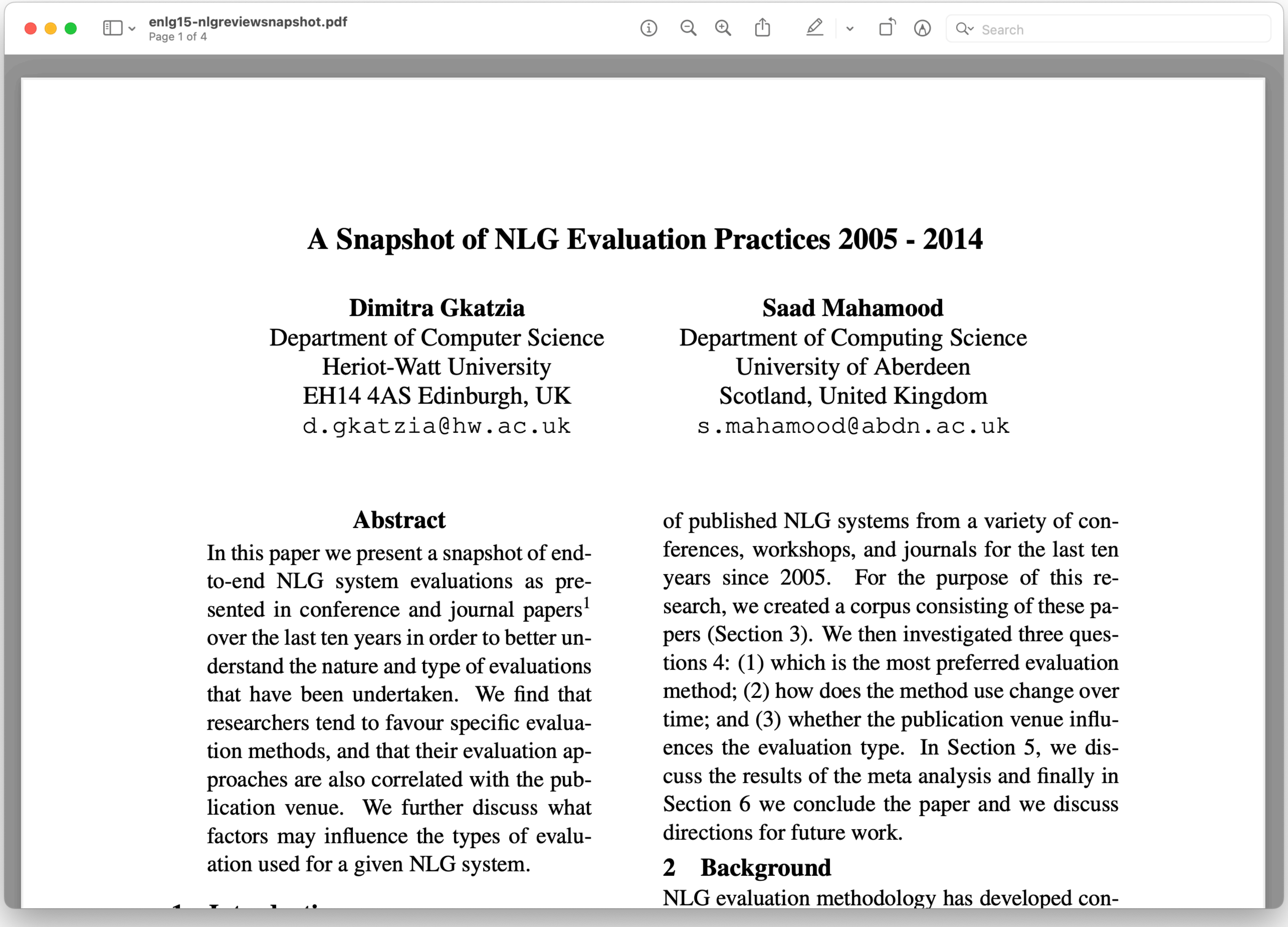
Our goal in this experiment was to shed light on when automatic metrics should be used in NLG. Given all the previously mentioned caveats, we cannot of course draw firm conclusions about this topic. But we can make some suggestions.

First of all, the automatic metrics we examined should not be used to predict human judgments of content quality; none of them had a significant correlation with human accuracy judgments, even when statistical significance is calculated in a less-than-conservative fashion.

Second of all, even when evaluating linguistic quality, current automatic metrics should be used with caution, as a supplement rather than a replacement for human evaluation; similar comments have been made about the use of automatic metrics in

My Journey into NLG Evaluation

- ~2010: First aware of issues in NLG evaluation, due to Reiter's & Belz's paper on investigating the validity of evaluations [Reiter & Belz, 2009]
- 2015: Worked with Dimitra Gkatzia on generating a snapshot of NLG evaluation practices for the last 10 years [Gkatzia & Mahamood, 2015].



A Snapshot of NLG Evaluation Practices 2005 - 2014

Dimitra Gkatzia
Department of Computer Science
Heriot-Watt University
EH14 4AS Edinburgh, UK
d.gkatzia@hw.ac.uk

Saad Mahamood
Department of Computing Science
University of Aberdeen
Scotland, United Kingdom
s.mahamood@abdn.ac.uk

Abstract

In this paper we present a snapshot of end-to-end NLG system evaluations as presented in conference and journal papers[1] over the last ten years in order to better understand the nature and type of evaluations that have been undertaken. We find that researchers tend to favour specific evaluation methods, and that their evaluation approaches are also correlated with the publication venue. We further discuss what factors may influence the types of evaluation used for a given NLG system.

1 Introduction

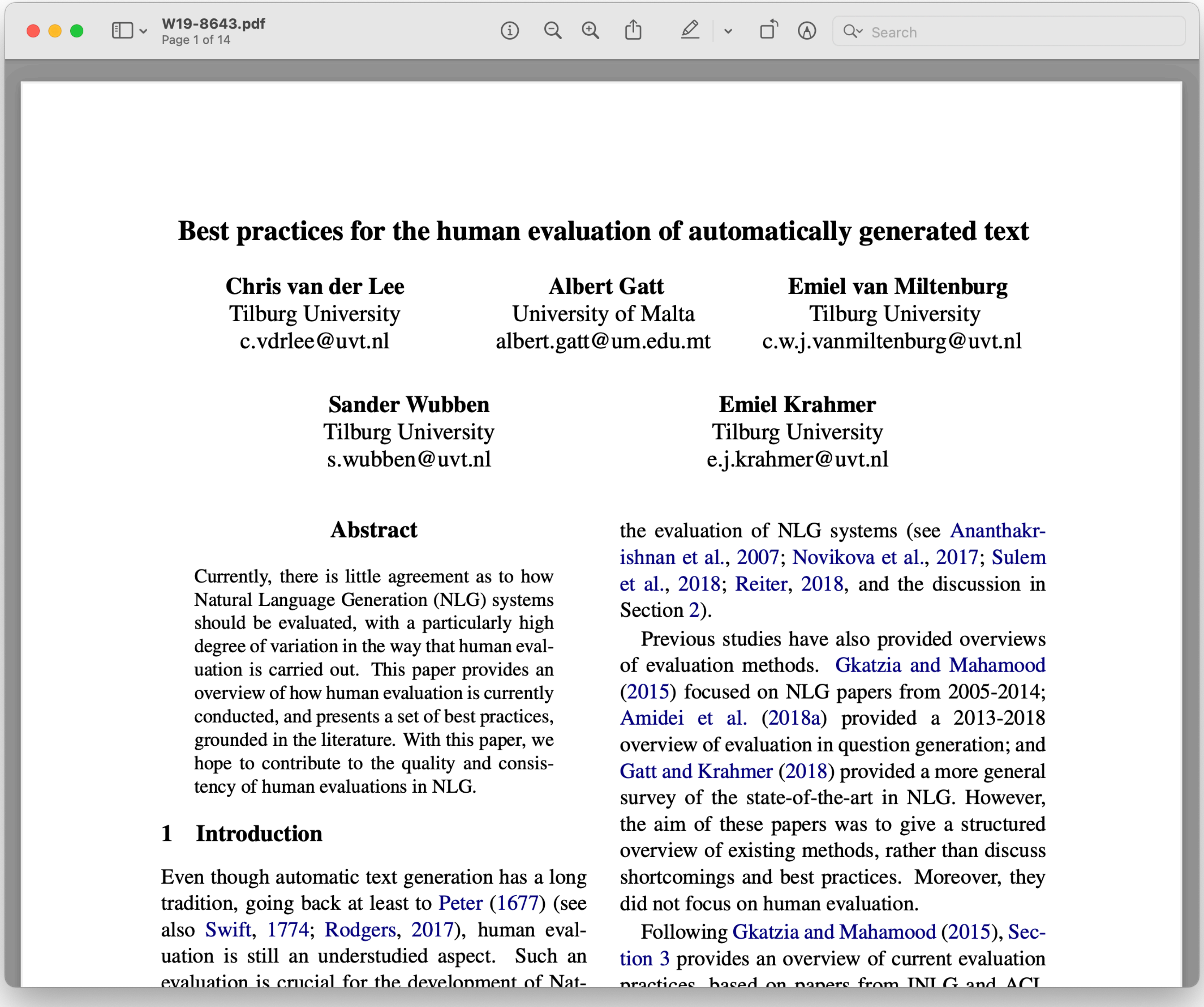
of published NLG systems from a variety of conferences, workshops, and journals for the last ten years since 2005. For the purpose of this research, we created a corpus consisting of these papers (Section 3). We then investigated three questions 4: (1) which is the most preferred evaluation method; (2) how does the method use change over time; and (3) whether the publication venue influences the evaluation type. In Section 5, we discuss the results of the meta analysis and finally in Section 6 we conclude the paper and we discuss directions for future work.

2 Background

NLG evaluation methodology has developed con-

My Journey into NLG Evaluation

- ~2010: First aware of issues in NLG evaluation, due to Reiter's & Belz's paper on investigating the validity of evaluations [Reiter & Belz, 2009]
- 2015: Worked with Dimitra Gkatzia on generating a snapshot of NLG evaluation practices for the last 10 years [Gkatzia & Mahamood, 2015].
- 2019: At INLG 2019, presentation on how Human Evaluations are conducted in NLG and their problems [van Der Lee et al., 2019].



Best practices for the human evaluation of automatically generated text

Chris van der Lee
Tilburg University
c.vdrlee@uvt.nl

Albert Gatt
University of Malta
albert.gatt@um.edu.mt

Emiel van Miltenburg
Tilburg University
c.w.j.vanmiltenburg@uvt.nl

Sander Wubben
Tilburg University
s.wubben@uvt.nl

Emiel Krahmer
Tilburg University
e.j.krahmer@uvt.nl

Abstract

Currently, there is little agreement as to how Natural Language Generation (NLG) systems should be evaluated, with a particularly high degree of variation in the way that human evaluation is carried out. This paper provides an overview of how human evaluation is currently conducted, and presents a set of best practices, grounded in the literature. With this paper, we hope to contribute to the quality and consistency of human evaluations in NLG.

1 Introduction

Even though automatic text generation has a long tradition, going back at least to Peter (1677) (see also Swift, 1774; Rodgers, 2017), human evaluation is still an understudied aspect. Such an evaluation is crucial for the development of Natthe evaluation of NLG systems (see Ananthakrishnan et al., 2007; Novikova et al., 2017; Sulem et al., 2018; Reiter, 2018, and the discussion in Section 2).

Previous studies have also provided overviews of evaluation methods. Gkatzia and Mahamood (2015) focused on NLG papers from 2005-2014; Amidei et al. (2018a) provided a 2013-2018 overview of evaluation in question generation; and Gatt and Krahmer (2018) provided a more general survey of the state-of-the-art in NLG. However, the aim of these papers was to give a structured overview of existing methods, rather than discuss shortcomings and best practices. Moreover, they did not focus on human evaluation.

Following Gkatzia and Mahamood (2015), Section 3 provides an overview of current evaluation practices, based on papers from INLG and ACL

My Journey into NLG Evaluation

- ~2010: First aware of issues in NLG evaluation, due to Reiter's & Belz's paper on investigating the validity of evaluations [Reiter & Belz, 2009]
- 2015: Worked with Dimitra Gkatzia on generating a snapshot of NLG evaluation practices for the last 10 years [Gkatzia & Mahamood, 2015].
- 2019: At INLG 2019, presentation on how Human Evaluations are conducted in NLG and their problems [van Der Lee et al., 2019].
- 2020: Collaborated with multiple researchers on an extensive 20-year overview of how NLG human evaluations are conducted [Howcroft et al., 2020].

My Journey into NLG Evaluation

- 2021: Worked on several evaluation related research initiatives:
  - Explored Commonsense human NLG evaluations [Clinciu et al., 2021]
  - Underreporting of errors in NLG output [Miltenberg et al, 2021]
  - Reproducing an earlier NLG experiment [Mahamood, 2021]
  - Automatic construction of evaluation test sets [Mille et al, 2021]

2. Overview: What are the issues?

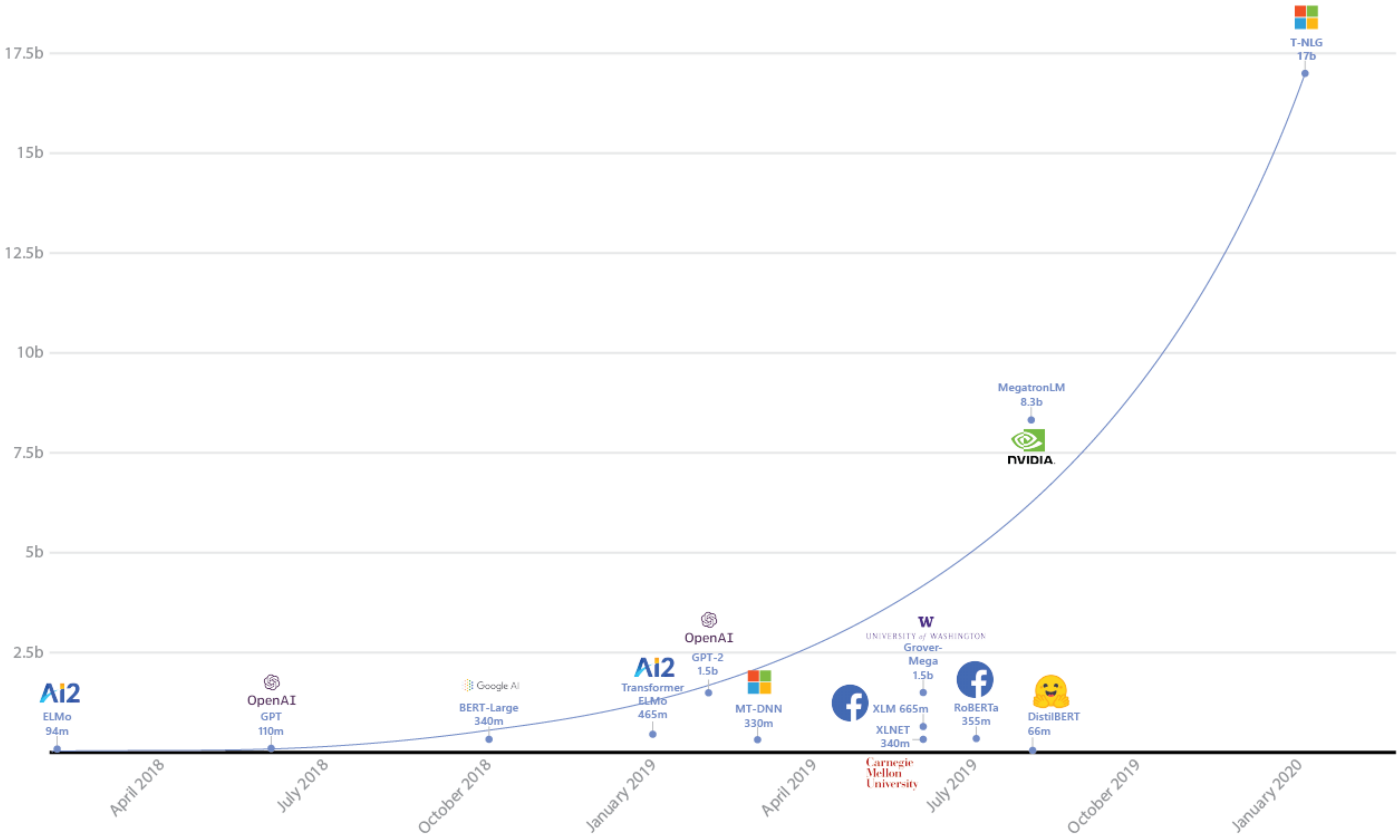
Overview

- To understand how well a NLG system or model performs we need to evaluate it's performance.
- There are two types of evaluation:
  - *Intrinsic evaluations* - e.g. How fluent is this text?
  - *Extrinsic evaluations* - e.g. How well does this generated report allow doctors make correct care decisions?
- Intrinsic evaluations can be performed using automated metrics or using human participants.
- Most evaluations performed in NLG research are intrinsic in nature [Gkatzia & Mahamood, 2015] and are automatic [Howcroft et al., 2020].

Automated Intrinsic Evaluations

Automatic Intrinsic evaluations have several problems…

- Most automated measures only measure similarity between model output and reference corpora [Gehrmann et al., 2022].
- Evaluations using metrics such as BLEU, ROUGE, etc. correlate poorly with human judgments [Reiter & Belz, 2009, Reiter, 2018] and small changes may not be statistically significant [Mathur, 2020].
- Most publications only use a single metric to demonstrate improvements over prior systems [Gehrmann et al., 2022].
- 100% of papers introducing new summarisation models at *CL conferences in 2021 use ROUGE and 69% use only ROUGE [Gehrmann et al., 2022].


17.5b
15b
12.5b
10b
7.5b
5b
2.5b
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020
AI2
ELMo
94m
OpenAI
GPT
110m
Google AI
BERT-Large
340m
AI2
Transformer
ELMo
465m
OpenAI
GPT-2
1.5b
MT-DNN
330m
XLM 665m
XLNET
340m
Carnegie
Mellon
University
W
UNIVERSITY of WASHINGTON
Grover-
Mega
1.5b
RoBERTa
355m
DistilBERT
66m
MegatronLM
8.3b
NVIDIA
T-NLG
17b

Neural NLG and Evaluations

However, neural NLG evaluations poses additional challenges:

- Factual accuracy and hallucinations [Thomson and Reiter, 2020].
- Ethical challenges. Large neural language models can reinforce discriminatory behaviour such as sexist gender roles, racists language, etc. [Bender et al., 2021].
- Therefore, there is a need to understand better how robust given neural models are and how well they perform under variety of datasets.
- Only by performing a multi-dimensional evaluations can we evaluate several aspects of a generated text's quality [Gehrmann et al., 2022].

Human NLG Evaluations

Additionally, human evaluations have their own problems…

- Human evaluations, whilst considered more reliable than automatic, have issues with extreme diversity in the approaches used and fundamental gaps in details being reported [Howcroft et al., 2020].
- Evaluations do not consistently name their quality criterion and definitions.

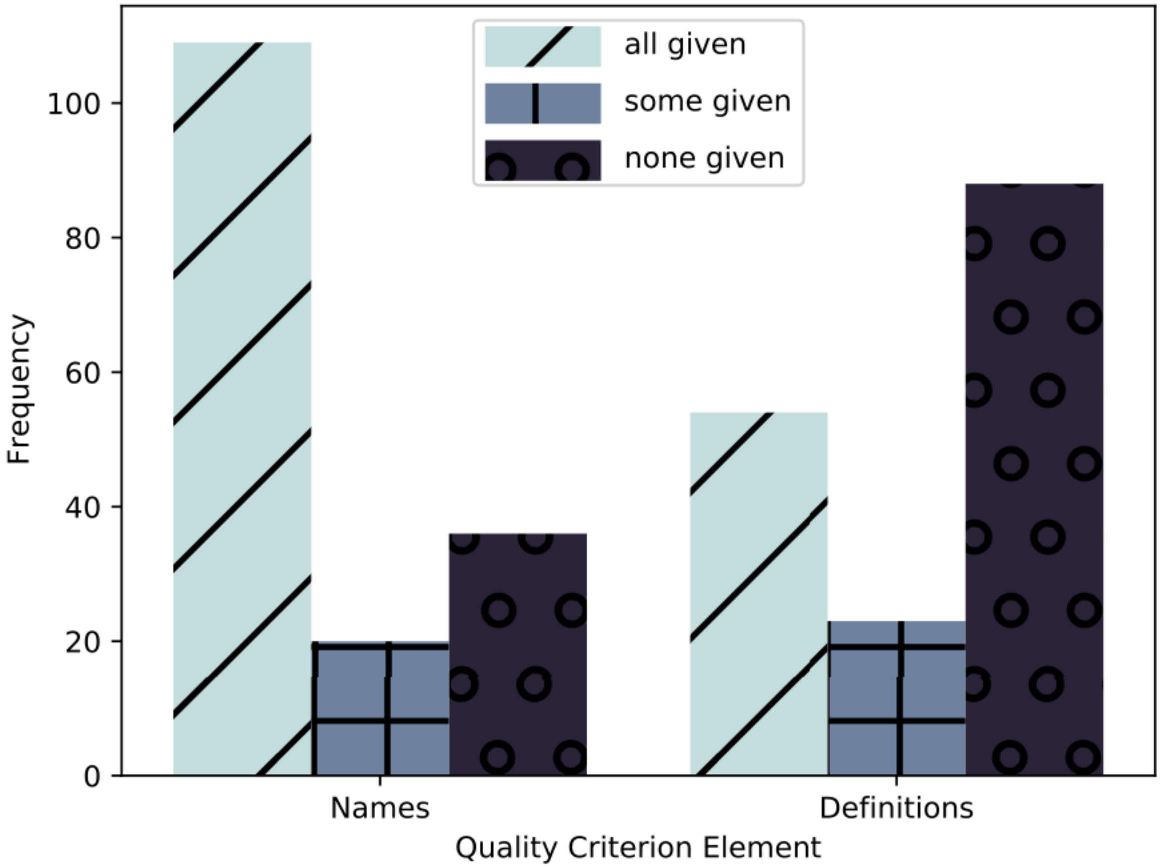


Figure 2: How many papers explicitly name and define all, some, or none of the quality criteria they evaluate.

Human NLG Evaluations

Additionally, human evaluations have their own problems…

- Human evaluations, whilst considered more reliable than automatic, have issues with extreme diversity in the approaches used and fundamental gaps in details being reported [Howcroft et al., 2020].
- Evaluations do not consistently name their quality criterion and definitions.
- There is uncertainty of what is being measured:
  - Howcroft et al., found generated text was evaluated on 204 dimensions of quality, which mapped to 71 distinct criteria.

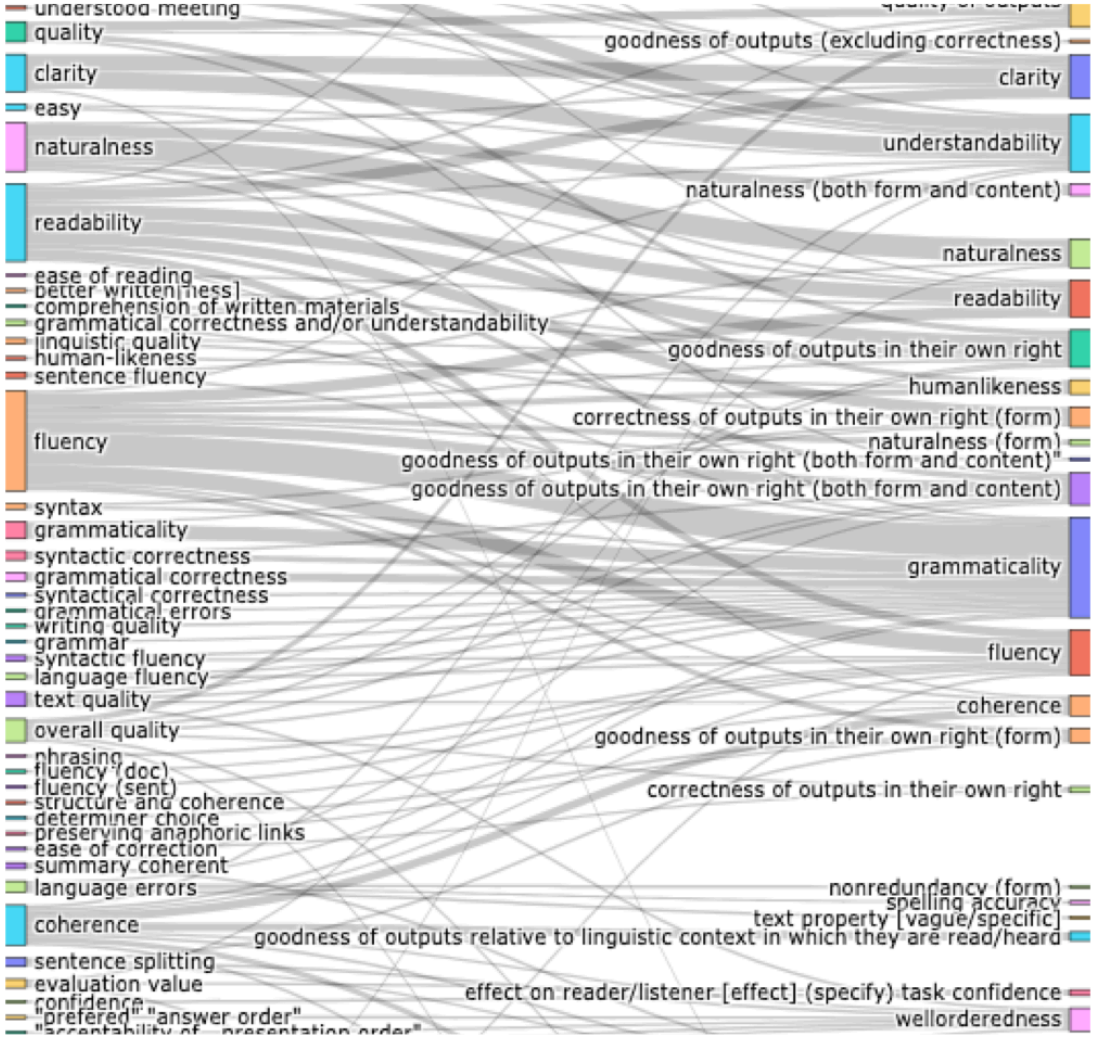

understood meeting
quality
clarity
easy
naturalness
readability
ease of reading
better written[ness]
comprehension of written materials
grammatical correctness and/or understandability
linguistic quality
human-likeness
sentence fluency
fluency
syntax
grammaticality
syntactic correctness
grammatical correctness
syntactical correctness
grammatical errors
writing quality
grammar
syntactic fluency
language fluency
text quality
overall quality
phrasing
fluency (doc)
fluency (sent)
structure and coherence
determiner choice
preserving anaphoric links
ease of correction
summary coherent
language errors
coherence
sentence splitting
evaluation value
confidence
"prefered" "answer order"
"acceptability of ... presentation order"
quality of outputs
goodness of outputs (excluding correctness)
clarity
understandability
naturalness (both form and content)
naturalness
readability
goodness of outputs in their own right
humanlikeness
correctness of outputs in their own right (form)
naturalness (form)
goodness of outputs in their own right (both form and content)"
goodness of outputs in their own right (both form and content)
grammaticality
fluency
coherence
goodness of outputs in their own right (form)
correctness of outputs in their own right
nonredundancy (form)
spelling accuracy
text property [vague/specific]
goodness of outputs relative to linguistic context in which they are read/heard
effect on reader/listener [effect] (specify) task confidence
wellorderedness

Figure 3: Part of Sankey diagram of evaluation criteria names from NLG papers between 2000 & 2019 (left) mapped to normalised criteria names representing our assessment of what was actually measured (right).

Reproducibility

Finally, can we even reproduce the results…

- There is a need to ensure that presented results are sound and reliable. This means they need to be reproducible.
- NeurIPS in 2019 introduced a machine learning reproducibility checklist for submissions [Pineau et al., 2020].
- Growing interest in trying to reproduce human evaluations within NLP.
- However, recent work look reproducibility in NLP found signifiant issues [Belz et al., 2021]:
  - Only a minority of systems reproducing previously reported scores.

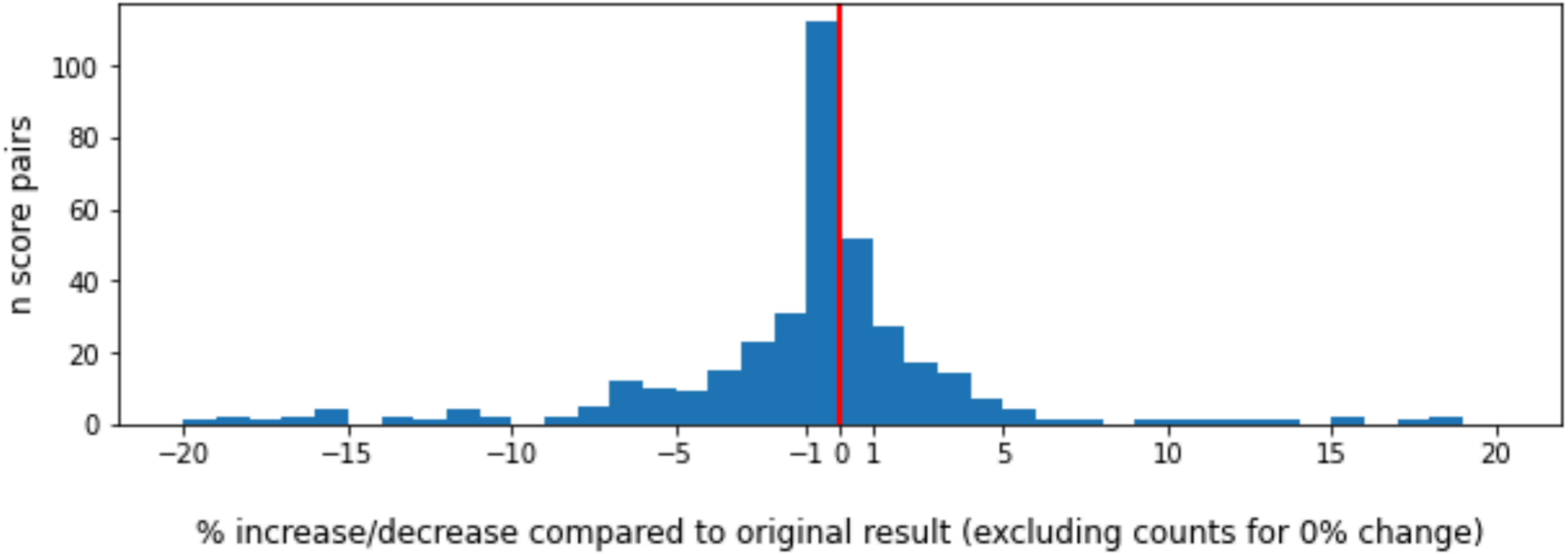


Figure 2: Histogram of percentage differences between original and reproduction scores (bin width = 1; clipped to range -20..20).

Reproducibility

Finally, can we even reproduce the results…

- There is a need to ensure that presented results are sound and reliable. This means they need to be reproducible.
- NeurIPS in 2019 introduced a machine learning reproducibility checklist for submissions [Pineau et al., 2020].
- Growing interest in trying to reproduce human evaluations within NLP.
- However, recent work look reproducibility in NLP found signifiant issues [Belz et al., 2021]:
    - Only a minority of systems reproducing previously reported scores.
    - Systems not working due to non-functional code or resource limits.

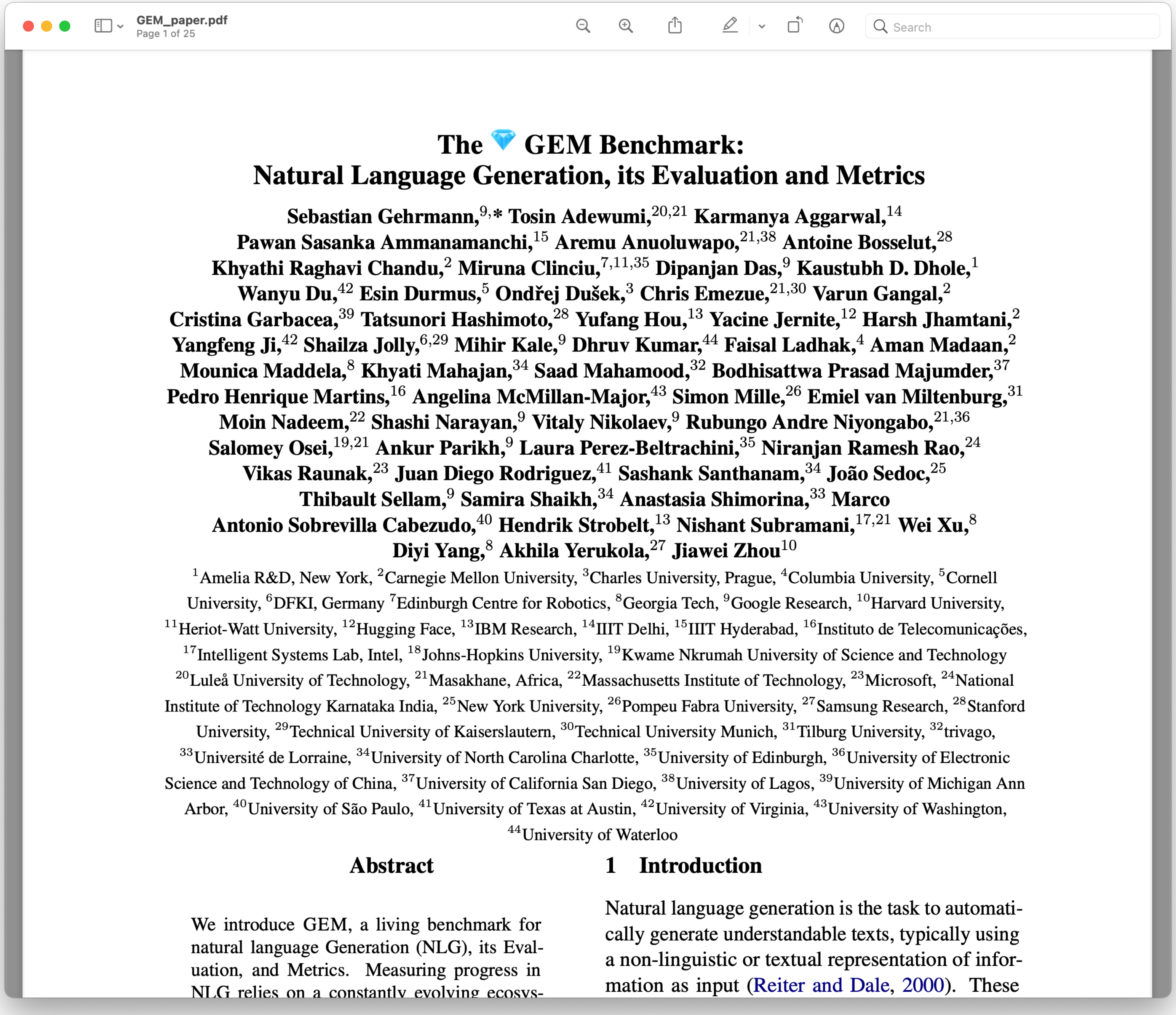
3. Neural NLG Evaluation Efforts

3.1. The GEM Project

The GEM Project

What the GEM project is:

- The Generation, Evaluation, and Metrics (GEM) project is a living benchmark that aims to evaluate in-depth models.
- The GEM project is a large collaborative research endeavour with collaborators from multiple continents and institutions.



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]
Khyathi Raghavi Chandu,[2] **Miruna Clinciu,**[7,11,35] **Dipanjan Das,**[9] **Kaustubh D. Dhole,**[1]
Wanyu Du,[42] **Esin Durmus,**[5] **Ondřej Dušek,**[3] **Chris Emezue,**[21,30] **Varun Gangal,**[2]
Cristina Garbacea,[39] **Tatsunori Hashimoto,**[28] **Yufang Hou,**[13] **Yacine Jernite,**[12] **Harsh Jhamtani,**[2]
Yangfeng Ji,[42] **Shailza Jolly,**[6,29] **Mihir Kale,**[9] **Dhruv Kumar,**[44] **Faisal Ladhak,**[4] **Aman Madaan,**[2]
Mounica Maddela,[8] **Khyati Mahajan,**[34] **Saad Mahamood,**[32] **Bodhisattwa Prasad Majumder,**[37]
Pedro Henrique Martins,[16] **Angelina McMillan-Major,**[43] **Simon Mille,**[26] **Emiel van Miltenburg,**[31]
Moin Nadeem,[22] **Shashi Narayan,**[9] **Vitaly Nikolaev,**[9] **Rubungo Andre Niyongabo,**[21,36]
Salomey Osei,[19,21] **Ankur Parikh,**[9] **Laura Perez-Beltrachini,**[35] **Niranjan Ramesh Rao,**[24]
Vikas Raunak,[23] **Juan Diego Rodriguez,**[41] **Sashank Santhanam,**[34] **João Sedoc,**[25]
Thibault Sellam,[9] **Samira Shaikh,**[34] **Anastasia Shimorina,**[33] **Marco Antonio Sobrevilla Cabezudo,**[40] **Hendrik Strobelt,**[13] **Nishant Subramani,**[17,21] **Wei Xu,**[8]
Diyi Yang,[8] **Akhila Yerukola,**[27] **Jiawei Zhou**[10]

[1]Amelia R&D, New York, [2]Carnegie Mellon University, [3]Charles University, Prague, [4]Columbia University, [5]Cornell University, [6]DFKI, Germany [7]Edinburgh Centre for Robotics, [8]Georgia Tech, [9]Google Research, [10]Harvard University, [11]Heriot-Watt University, [12]Hugging Face, [13]IBM Research, [14]IIIT Delhi, [15]IIIT Hyderabad, [16]Instituto de Telecomunicações, [17]Intelligent Systems Lab, Intel, [18]Johns-Hopkins University, [19]Kwame Nkrumah University of Science and Technology [20]Luleå University of Technology, [21]Masakhane, Africa, [22]Massachusetts Institute of Technology, [23]Microsoft, [24]National Institute of Technology Karnataka India, [25]New York University, [26]Pompeu Fabra University, [27]Samsung Research, [28]Stanford University, [29]Technical University of Kaiserslautern, [30]Technical University Munich, [31]Tilburg University, [32]trivago, [33]Université de Lorraine, [34]University of North Carolina Charlotte, [35]University of Edinburgh, [36]University of Electronic Science and Technology of China, [37]University of California San Diego, [38]University of Lagos, [39]University of Michigan Ann Arbor, [40]University of São Paulo, [41]University of Texas at Austin, [42]University of Virginia, [43]University of Washington, [44]University of Waterloo

Abstract

We introduce GEM, a living benchmark for natural language Generation (NLG), its Evaluation, and Metrics. Measuring progress in NLG relies on a constantly evolving ecosys-

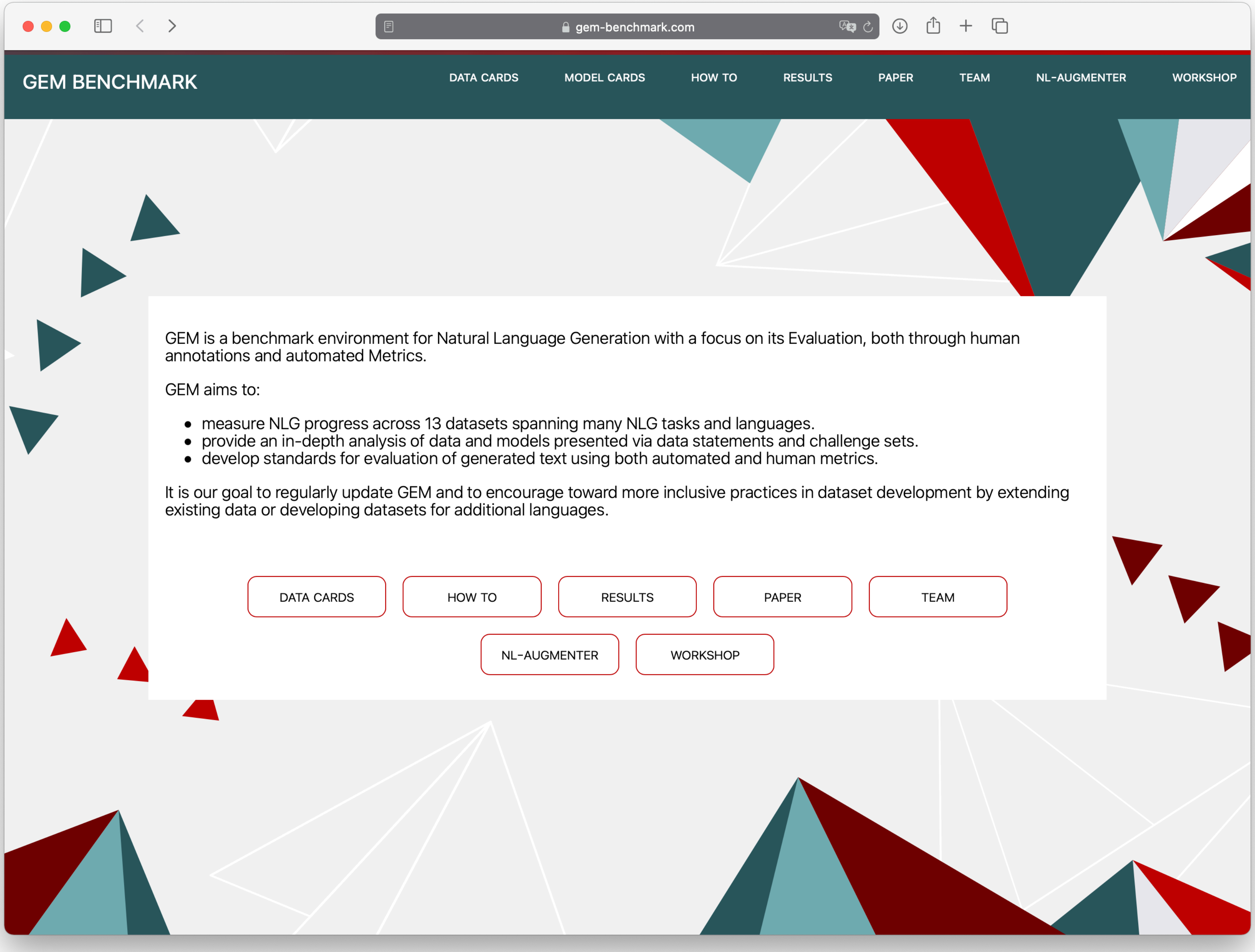
1 Introduction

Natural language generation is the task to automatically generate understandable texts, typically using a non-linguistic or textual representation of information as input (Reiter and Dale, 2000). These

The GEM Project

What the GEM project is:

- The Generation, Evaluation, and Metrics (GEM) project is a living benchmark that aims to evaluate in-depth models.
- The GEM project is a large collaborative research endeavour with collaborators from multiple continents and institutions.
- The goal is to have an accurate representation of model performance, uncover shortcomings, and opportunities for progress.

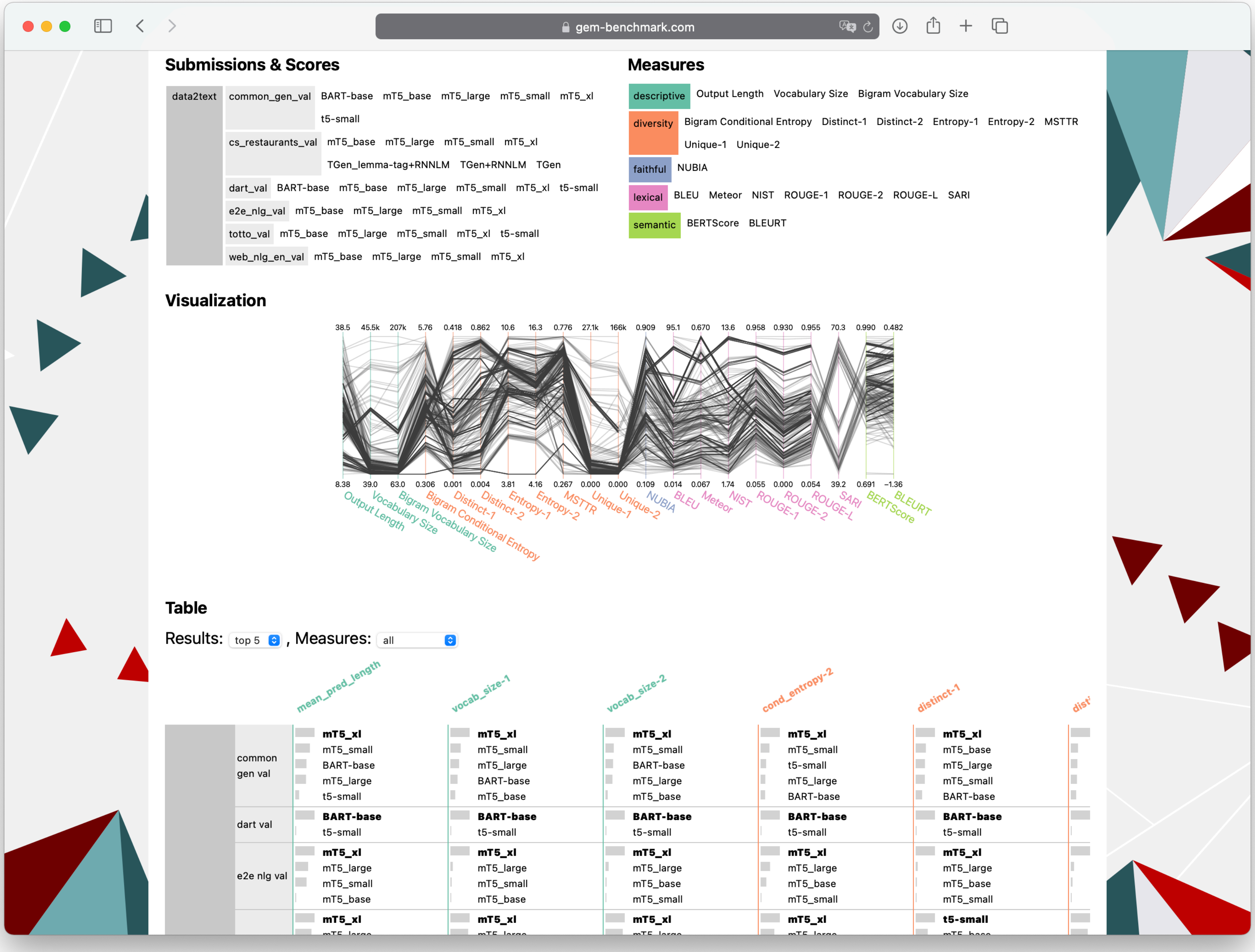


GEM is a benchmark environment for Natural Language Generation with a focus on its Evaluation, both through human annotations and automated Metrics.

GEM aims to:

- measure NLG progress across 13 datasets spanning many NLG tasks and languages.
- provide an in-depth analysis of data and models presented via data statements and challenge sets.
- develop standards for evaluation of generated text using both automated and human metrics.

It is our goal to regularly update GEM and to encourage toward more inclusive practices in dataset development by extending existing data or developing datasets for additional languages.



Submissions & Scores

data2text	common_gen_val	BART-base mT5_base mT5_large mT5_small mT5_xl t5-small
	cs_restaurants_val	mT5_base mT5_large mT5_small mT5_xl TGen_lemma-tag+RNNLM TGen+RNNLM TGen
	dart_val	BART-base mT5_base mT5_large mT5_small mT5_xl t5-small
	e2e_nlg_val	mT5_base mT5_large mT5_small mT5_xl
	totto_val	mT5_base mT5_large mT5_small mT5_xl t5-small
	web_nlg_en_val	mT5_base mT5_large mT5_small mT5_xl

Measures

descriptive	Output Length Vocabulary Size Bigram Vocabulary Size
diversity	Bigram Conditional Entropy Distinct-1 Distinct-2 Entropy-1 Entropy-2 MSTTR Unique-1 Unique-2
faithful	NUBIA
lexical	BLEU Meteor NIST ROUGE-1 ROUGE-2 ROUGE-L SARI
semantic	BERTScore BLEURT

Visualization

Table

Results: top 5 , Measures: all

		mean_pred_length	vocab_size-1	vocab_size-2	cond_entropy-2	distinct-1	dist
	common gen val	**mT5_xl** mT5_small BART-base mT5_large t5-small	**mT5_xl** mT5_small mT5_large BART-base mT5_base	**mT5_xl** mT5_small BART-base mT5_large mT5_base	**mT5_xl** mT5_small t5-small mT5_large BART-base	**mT5_xl** mT5_base mT5_large mT5_small BART-base	
	dart val	**BART-base** t5-small	**BART-base** t5-small	**BART-base** t5-small	**BART-base** t5-small	**BART-base** t5-small	
	e2e nlg val	**mT5_xl** mT5_large mT5_small mT5_base	**mT5_xl** mT5_large mT5_small mT5_base	**mT5_xl** mT5_large mT5_base mT5_small	**mT5_xl** mT5_large mT5_base mT5_small	**mT5_xl** mT5_large mT5_base mT5_small	
		mT5_xl mT5_large	**mT5_xl** mT5_large	**mT5_xl** mT5_large	**mT5_xl** mT5_large	**t5-small** mT5_base	

The GEM Project

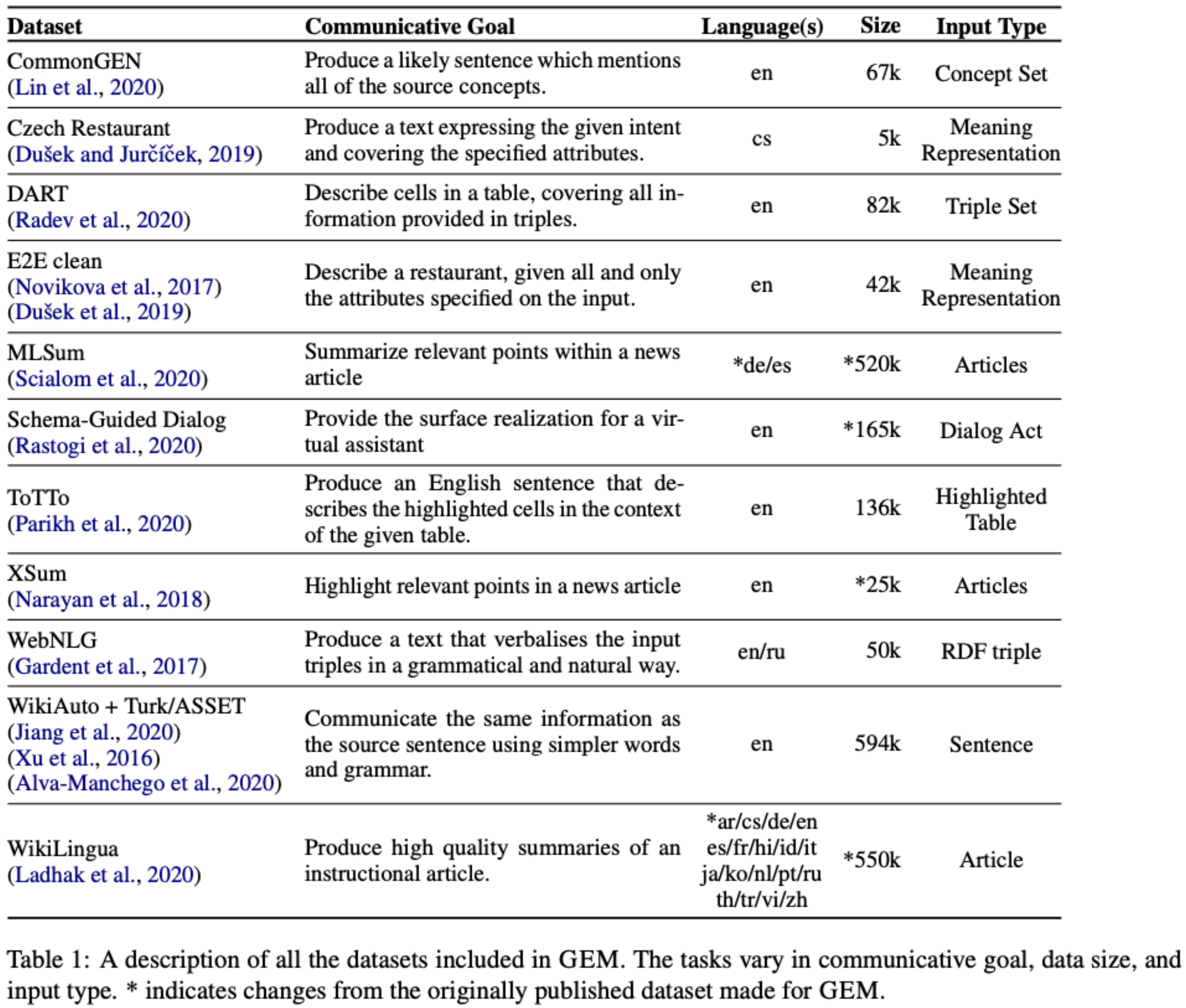
What the GEM project aims to achieve:

- Increasing multilingualism of NLG research.
    - Most benchmarks in NLG focuses exclusively in English.
- Providing a test bed for automated evaluations.
    - Allows for the latest advances into automated metrics to be test against a variety of NLG tasks such Data-to-Text, Summarisation, etc.
- Developing reproducible human standards.
    - Develop new standards for how human evaluations should be conducted, whilst incorporating lessons from related work e.g. WMT shared tasks.

The GEM Project

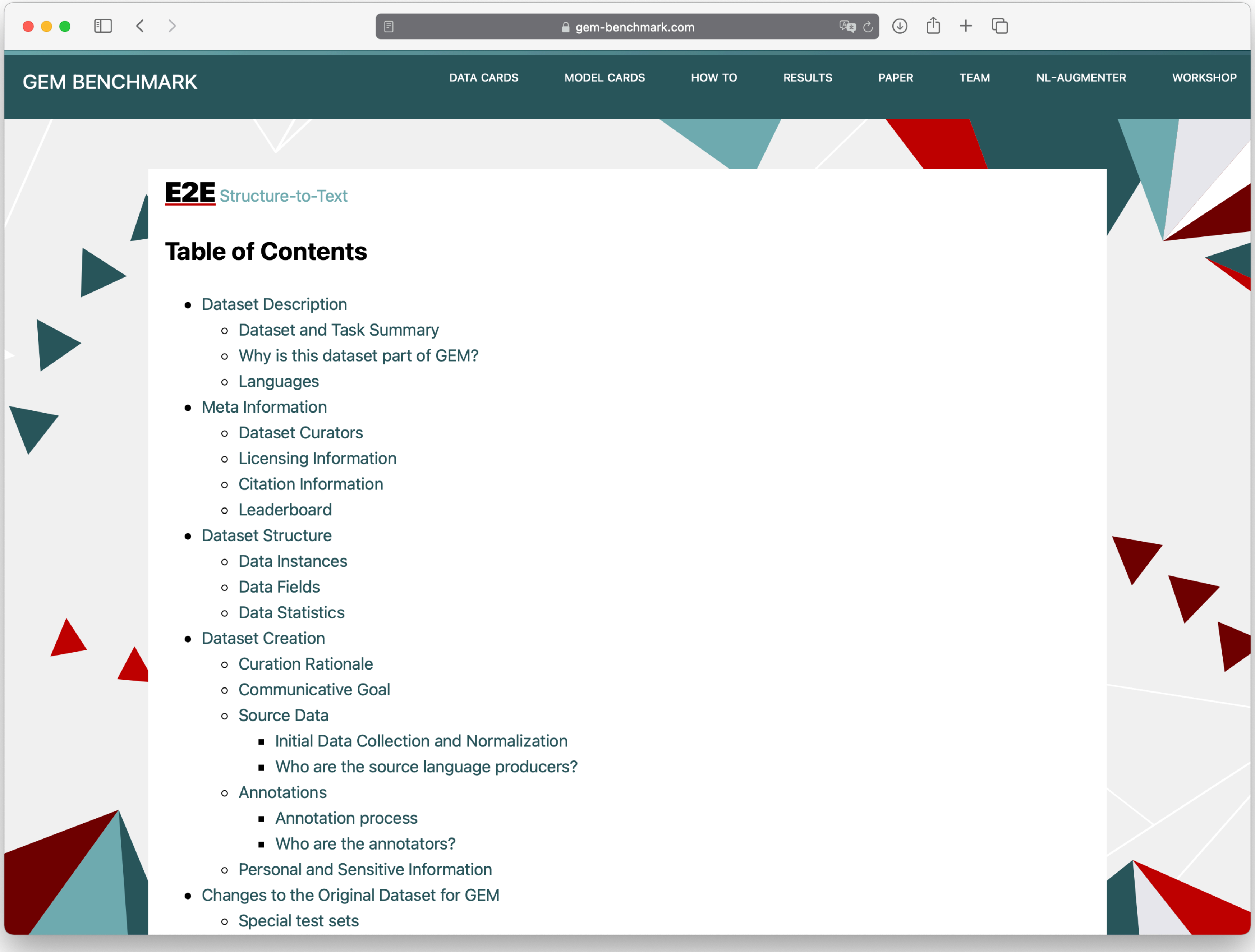
Selecting datasets for GEM:

- Choosing what datasets should be used is the most important part of a benchmark.
- Should be challenging to a variety of models, but still possible to evaluate models trained on them.
- We focused on datasets that:
  - Focus on diverse high-level tasks over a single high-level task.
  - Clean datasets to avoid conflating model mistakes and learned noise.
  - Mix of high- and low-resource datasets
  - Data with interesting test sets.
  - Focus not on the quality of current evaluation strategies.
  - Prefer multi-reference datasets, since those are shown robustness for automatic evaluation.



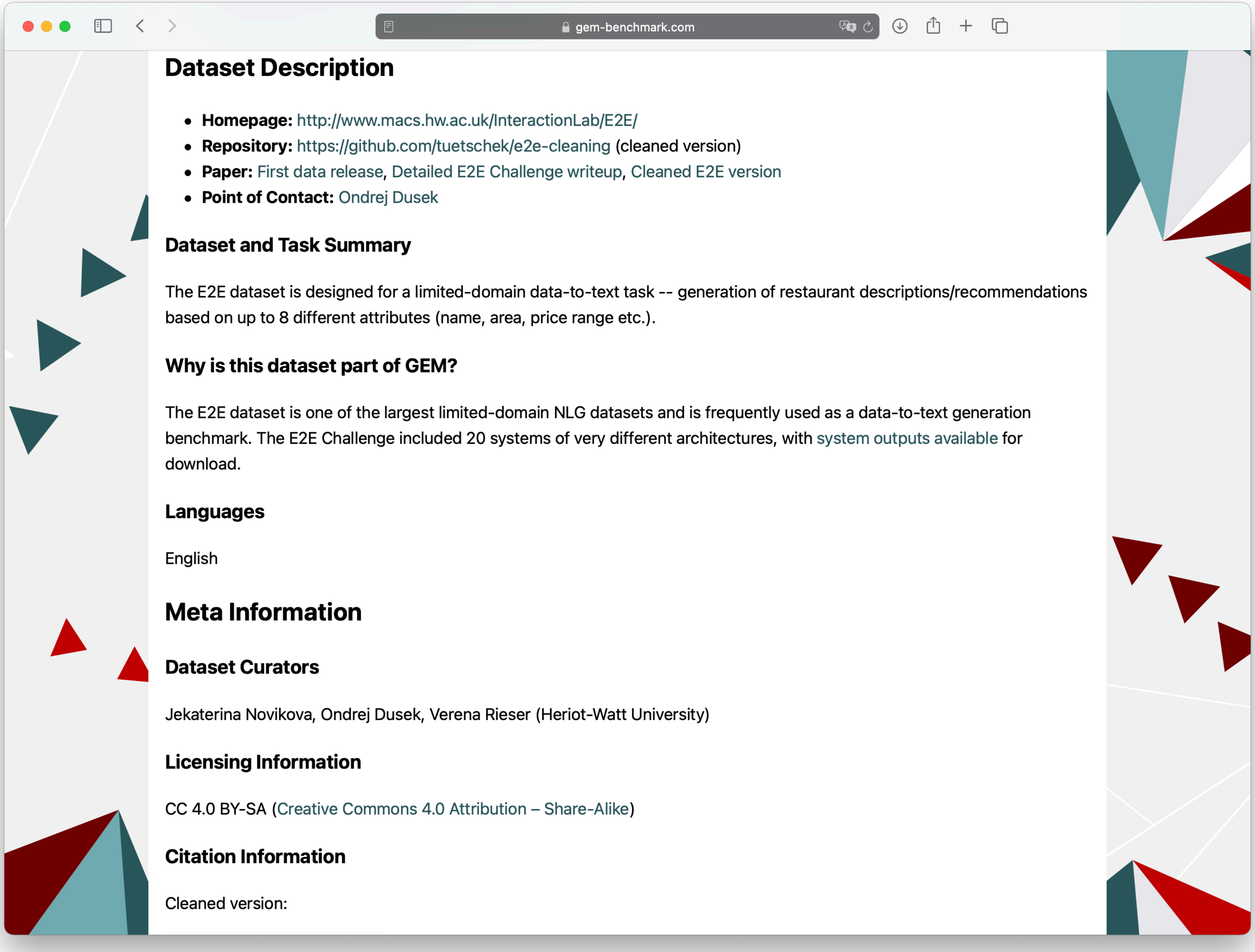
Dataset	Communicative Goal	Language(s)	Size	Input Type
CommonGEN (Lin et al., 2020)	Produce a likely sentence which mentions all of the source concepts.	en	67k	Concept Set
Czech Restaurant (Dušek and Jurčíček, 2019)	Produce a text expressing the given intent and covering the specified attributes.	cs	5k	Meaning Representation
DART (Radev et al., 2020)	Describe cells in a table, covering all information provided in triples.	en	82k	Triple Set
E2E clean (Novikova et al., 2017) (Dušek et al., 2019)	Describe a restaurant, given all and only the attributes specified on the input.	en	42k	Meaning Representation
MLSum (Scialom et al., 2020)	Summarize relevant points within a news article	*de/es	*520k	Articles
Schema-Guided Dialog (Rastogi et al., 2020)	Provide the surface realization for a virtual assistant	en	*165k	Dialog Act
ToTTo (Parikh et al., 2020)	Produce an English sentence that describes the highlighted cells in the context of the given table.	en	136k	Highlighted Table
XSum (Narayan et al., 2018)	Highlight relevant points in a news article	en	*25k	Articles
WebNLG (Gardent et al., 2017)	Produce a text that verbalises the input triples in a grammatical and natural way.	en/ru	50k	RDF triple
WikiAuto + Turk/ASSET (Jiang et al., 2020) (Xu et al., 2016) (Alva-Manchego et al., 2020)	Communicate the same information as the source sentence using simpler words and grammar.	en	594k	Sentence
WikiLingua (Ladhak et al., 2020)	Produce high quality summaries of an instructional article.	*ar/cs/de/en es/fr/hi/id/it ja/ko/nl/pt/ru th/tr/vi/zh	*550k	Article

Table 1: A description of all the datasets included in GEM. The tasks vary in communicative goal, data size, and input type. * indicates changes from the originally published dataset made for GEM.



E2E Structure-to-Text

Table of Contents



Dataset Description

- **Homepage:** http://www.macs.hw.ac.uk/InteractionLab/E2E/
- **Repository:** https://github.com/tuetschek/e2e-cleaning (cleaned version)
- **Paper:** First data release, Detailed E2E Challenge writeup, Cleaned E2E version
- **Point of Contact:** Ondrej Dusek

Dataset and Task Summary

The E2E dataset is designed for a limited-domain data-to-text task -- generation of restaurant descriptions/recommendations based on up to 8 different attributes (name, area, price range etc.).

Why is this dataset part of GEM?

The E2E dataset is one of the largest limited-domain NLG datasets and is frequently used as a data-to-text generation benchmark. The E2E Challenge included 20 systems of very different architectures, with system outputs available for download.

Languages

English

Meta Information

Dataset Curators

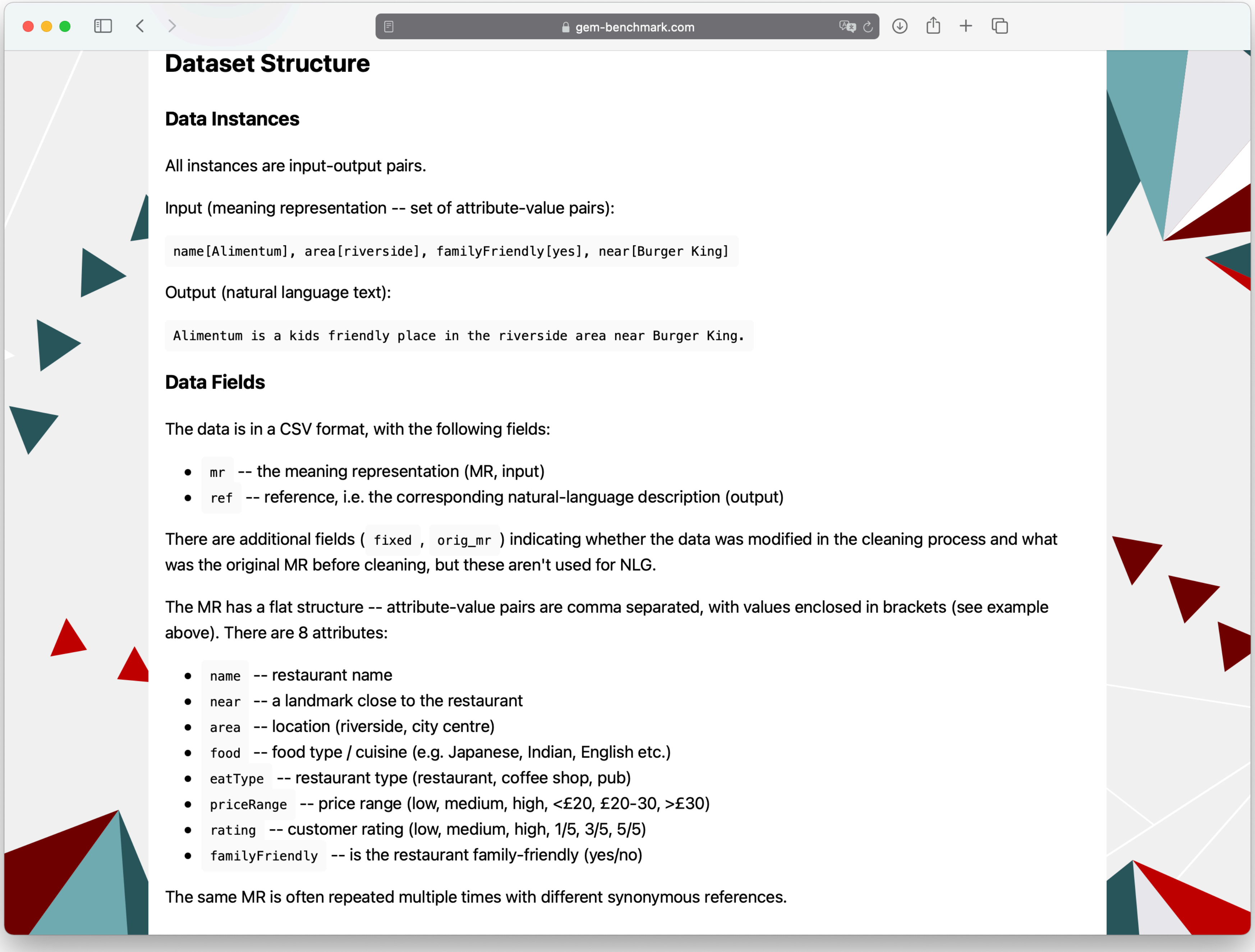
Jekaterina Novikova, Ondrej Dusek, Verena Rieser (Heriot-Watt University)

Licensing Information

CC 4.0 BY-SA (Creative Commons 4.0 Attribution – Share-Alike)

Citation Information

Cleaned version:



Dataset Structure

Data Instances

All instances are input-output pairs.

Input (meaning representation -- set of attribute-value pairs):

```
name[Alimentum], area[riverside], familyFriendly[yes], near[Burger King]
```

Output (natural language text):

```
Alimentum is a kids friendly place in the riverside area near Burger King.
```

Data Fields

The data is in a CSV format, with the following fields:

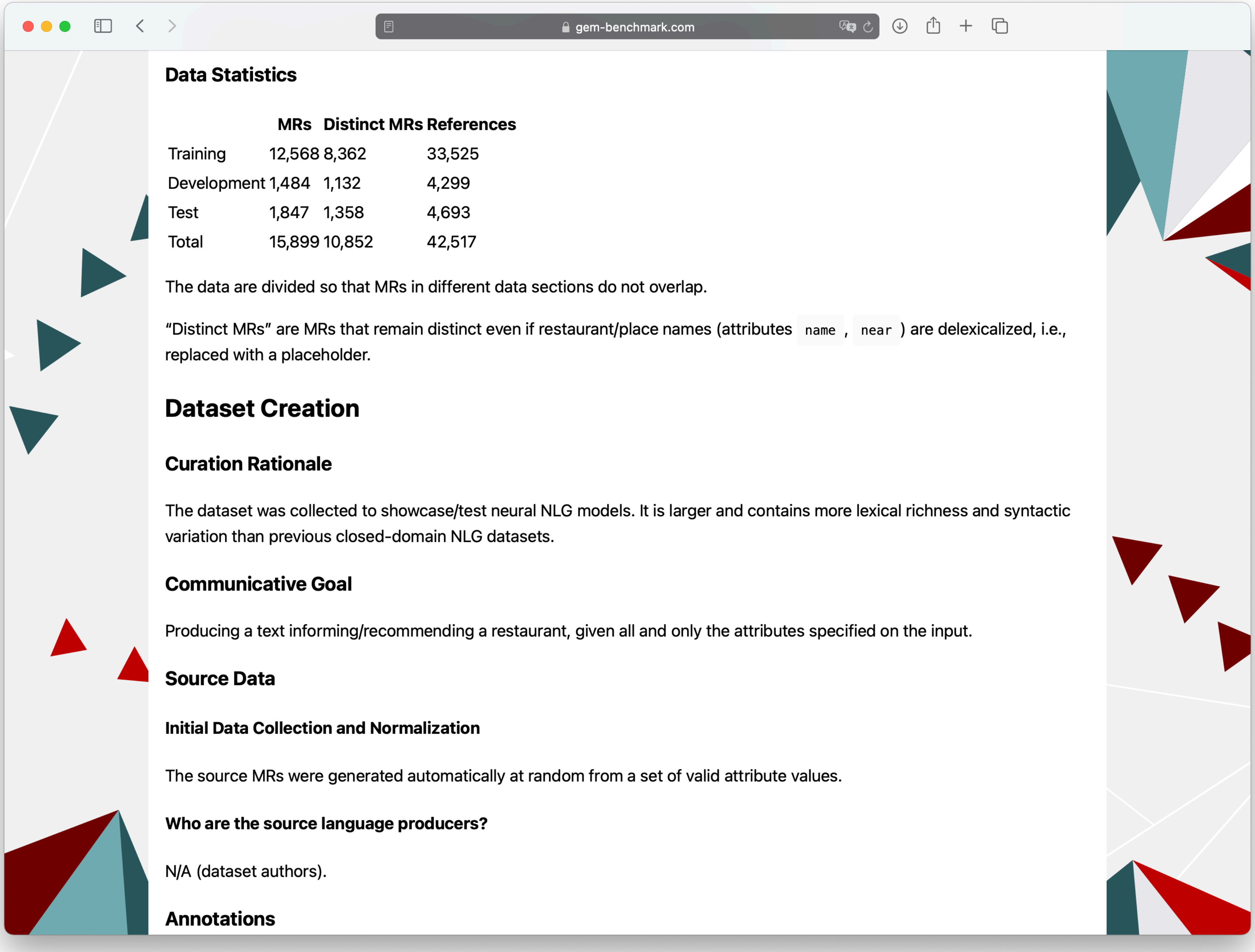
- `mr` -- the meaning representation (MR, input)
- `ref` -- reference, i.e. the corresponding natural-language description (output)

There are additional fields (`fixed` , `orig_mr`) indicating whether the data was modified in the cleaning process and what was the original MR before cleaning, but these aren't used for NLG.

The MR has a flat structure -- attribute-value pairs are comma separated, with values enclosed in brackets (see example above). There are 8 attributes:

- `name` -- restaurant name
- `near` -- a landmark close to the restaurant
- `area` -- location (riverside, city centre)
- `food` -- food type / cuisine (e.g. Japanese, Indian, English etc.)
- `eatType` -- restaurant type (restaurant, coffee shop, pub)
- `priceRange` -- price range (low, medium, high, <£20, £20-30, >£30)
- `rating` -- customer rating (low, medium, high, 1/5, 3/5, 5/5)
- `familyFriendly` -- is the restaurant family-friendly (yes/no)

The same MR is often repeated multiple times with different synonymous references.



Data Statistics

	MRs	Distinct MRs	References
Training	12,568	8,362	33,525
Development	1,484	1,132	4,299
Test	1,847	1,358	4,693
Total	15,899	10,852	42,517

The data are divided so that MRs in different data sections do not overlap.

“Distinct MRs” are MRs that remain distinct even if restaurant/place names (attributes `name` , `near`) are delexicalized, i.e., replaced with a placeholder.

Dataset Creation

Curation Rationale

The dataset was collected to showcase/test neural NLG models. It is larger and contains more lexical richness and syntactic variation than previous closed-domain NLG datasets.

Communicative Goal

Producing a text informing/recommending a restaurant, given all and only the attributes specified on the input.

Source Data

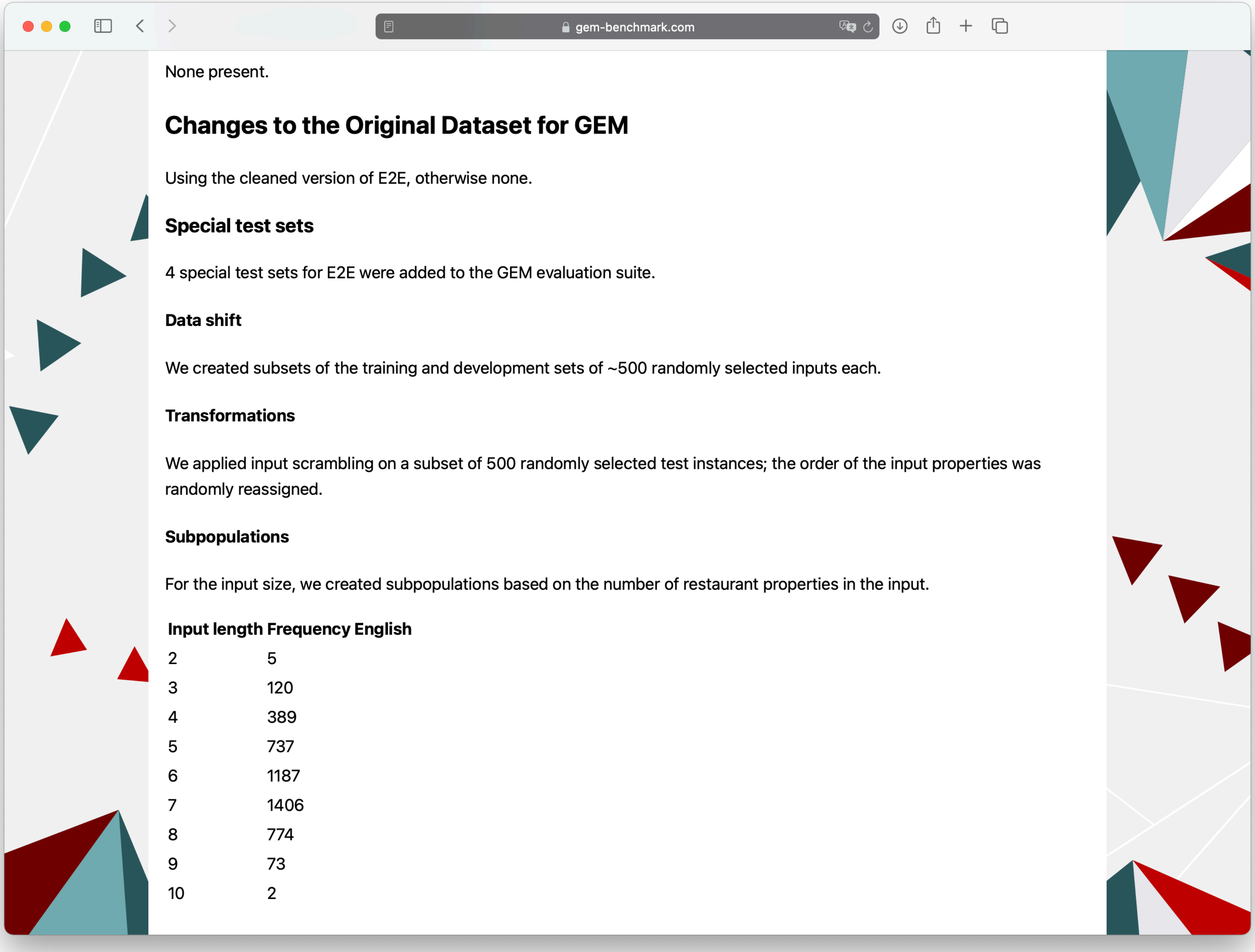
Initial Data Collection and Normalization

The source MRs were generated automatically at random from a set of valid attribute values.

Who are the source language producers?

N/A (dataset authors).

Annotations



None present.

Changes to the Original Dataset for GEM

Using the cleaned version of E2E, otherwise none.

Special test sets

4 special test sets for E2E were added to the GEM evaluation suite.

Data shift

We created subsets of the training and development sets of ~500 randomly selected inputs each.

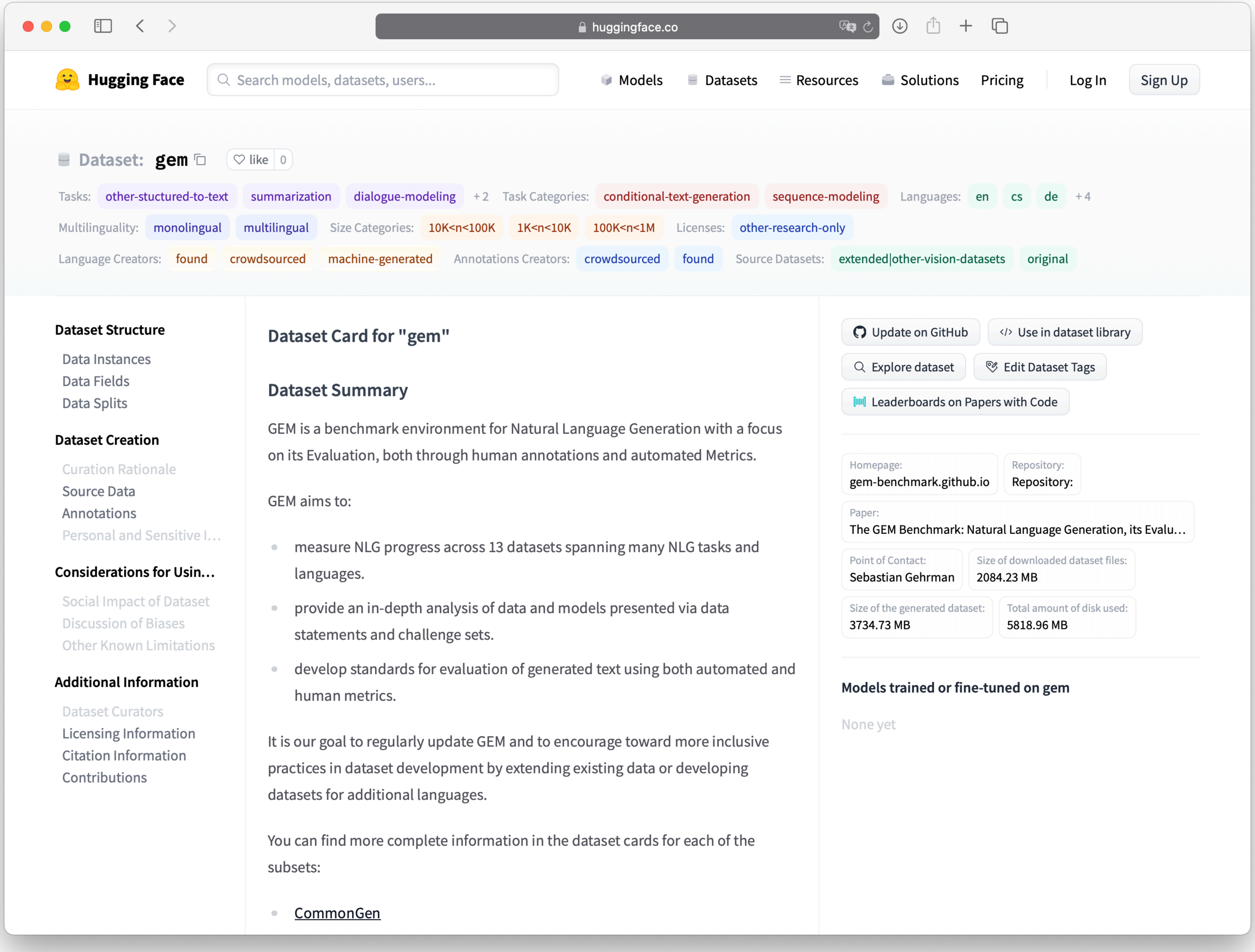
Transformations

We applied input scrambling on a subset of 500 randomly selected test instances; the order of the input properties was randomly reassigned.

Subpopulations

For the input size, we created subpopulations based on the number of restaurant properties in the input.

Input length	Frequency English
2	5
3	120
4	389
5	737
6	1187
7	1406
8	774
9	73
10	2



Dataset: **gem**

Tasks: other-stuctured-to-text summarization dialogue-modeling +2 Task Categories: conditional-text-generation sequence-modeling Languages: en cs de +4

Multilinguality: monolingual multilingual Size Categories: 10K<n<100K 1K<n<10K 100K<n<1M Licenses: other-research-only

Language Creators: found crowdsourced machine-generated Annotations Creators: crowdsourced found Source Datasets: extended|other-vision-datasets original

Dataset Structure

- Data Instances
- Data Fields
- Data Splits

Dataset Creation

- Curation Rationale
- Source Data
- Annotations
- Personal and Sensitive I...

Considerations for Usin...

- Social Impact of Dataset
- Discussion of Biases
- Other Known Limitations

Additional Information

- Dataset Curators
- Licensing Information
- Citation Information
- Contributions

Dataset Card for "gem"

Dataset Summary

GEM is a benchmark environment for Natural Language Generation with a focus on its Evaluation, both through human annotations and automated Metrics.

GEM aims to:

- measure NLG progress across 13 datasets spanning many NLG tasks and languages.
- provide an in-depth analysis of data and models presented via data statements and challenge sets.
- develop standards for evaluation of generated text using both automated and human metrics.

It is our goal to regularly update GEM and to encourage toward more inclusive practices in dataset development by extending existing data or developing datasets for additional languages.

You can find more complete information in the dataset cards for each of the subsets:

- CommonGen

Homepage: gem-benchmark.github.io

Repository: Repository:

Paper: The GEM Benchmark: Natural Language Generation, its Evalu...

Point of Contact: Sebastian Gehrman

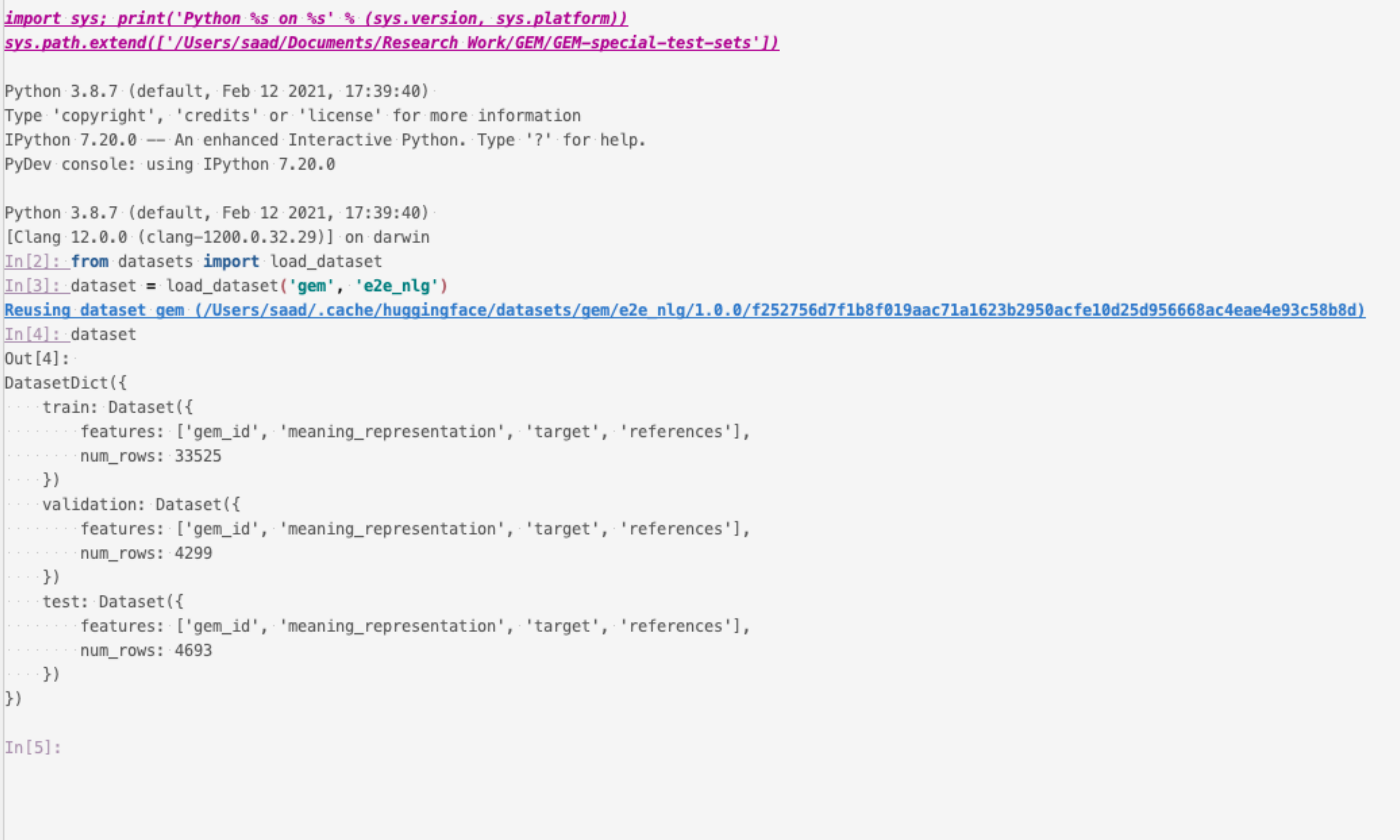
Size of downloaded dataset files: 2084.23 MB

Size of the generated dataset: 3734.73 MB

Total amount of disk used: 5818.96 MB

Models trained or fine-tuned on gem

None yet

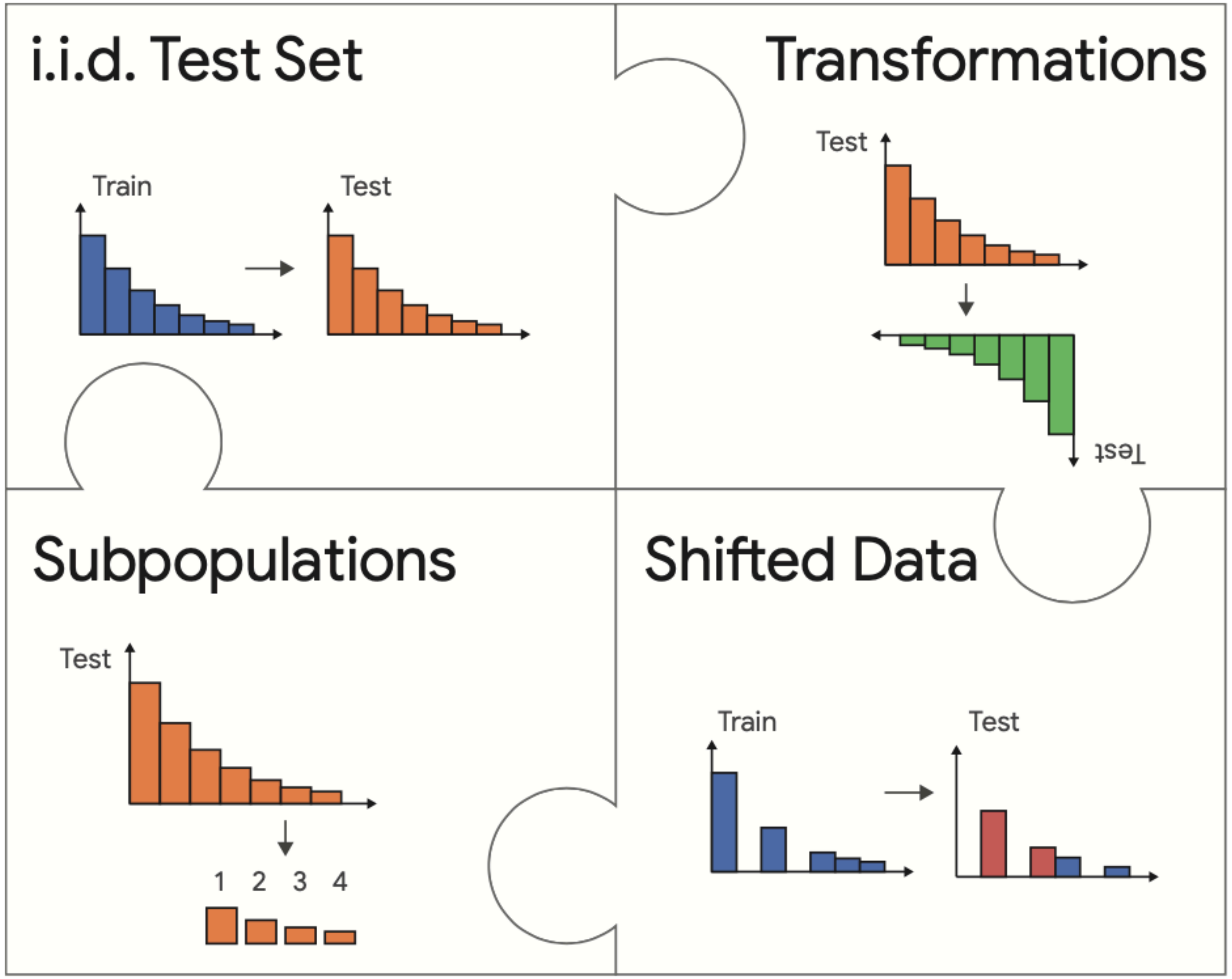


```
import sys; print('Python %s on %s' % (sys.version, sys.platform))
sys.path.extend(['/Users/saad/Documents/Research Work/GEM/GEM-special-test-sets'])

Python 3.8.7 (default, Feb 12 2021, 17:39:40) 
Type 'copyright', 'credits' or 'license' for more information
IPython 7.20.0 -- An enhanced Interactive Python. Type '?' for help.
PyDev console: using IPython 7.20.0

Python 3.8.7 (default, Feb 12 2021, 17:39:40) 
[Clang 12.0.0 (clang-1200.0.32.29)] on darwin
In[2]: from datasets import load_dataset
In[3]: dataset = load_dataset('gem', 'e2e_nlg')
Reusing dataset gem (/Users/saad/.cache/huggingface/datasets/gem/e2e_nlg/1.0.0/f252756d7f1b8f019aac71a1623b2950acfe10d25d956668ac4eae4e93c58b8d)
In[4]: dataset
Out[4]: 
DatasetDict({
    train: Dataset({
        features: ['gem_id', 'meaning_representation', 'target', 'references'],
        num_rows: 33525
    })
    validation: Dataset({
        features: ['gem_id', 'meaning_representation', 'target', 'references'],
        num_rows: 4299
    })
    test: Dataset({
        features: ['gem_id', 'meaning_representation', 'target', 'references'],
        num_rows: 4693
    })
})

In[5]: 
```

3.2. Generating Challenge Sets

Generating Challenge Sets

- In addition to using existing datasets we generated new special challenge sets.
- The purpose of these challenge sets is to go beyond evaluating models with just an independently and identical distributed test splits and expose how a model performs in the presence of challenging inputs.
- By altering existing datasets we can create new challenge sets to give us a better understanding of model robustness.



Subpopulations
Input Types [Dialog Acts, Topics]
Frequency [Complexity, Overlap]
Named Entity Features [Demographics]
Shape [Syntactic Structure, Properties]
Size [Length]
...

Transformations
Back-Translation [great -> toll -> fantastic]
Typos [English -> Eglish]
Punctuation [English. -> English]
Numerical Values [66 -> 79]
Scrambling [Cheap English -> English Cheap]
...

Shifted Data
Time-Shifted [COVID]
Data Samples [Train/Validation Examples]
...

Figure 1: Illustration of the types of evaluation suites that can be constructed from a given dataset.

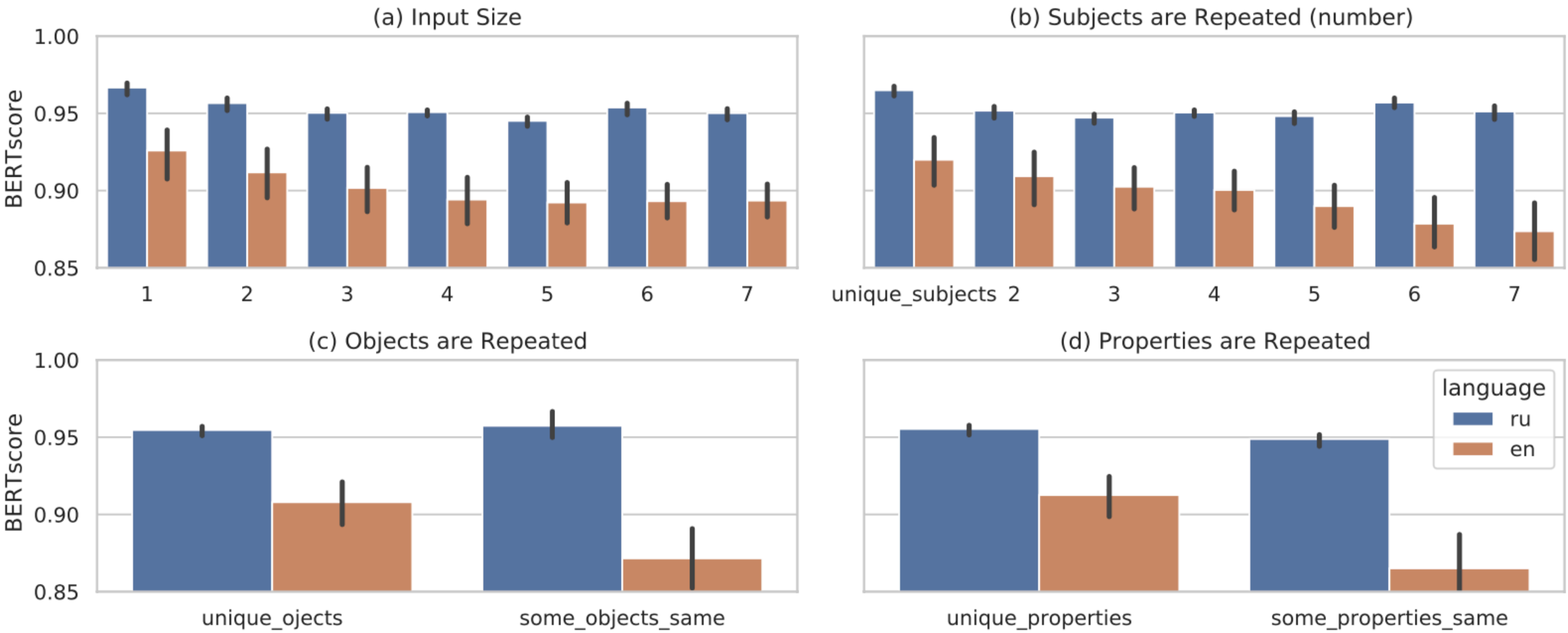
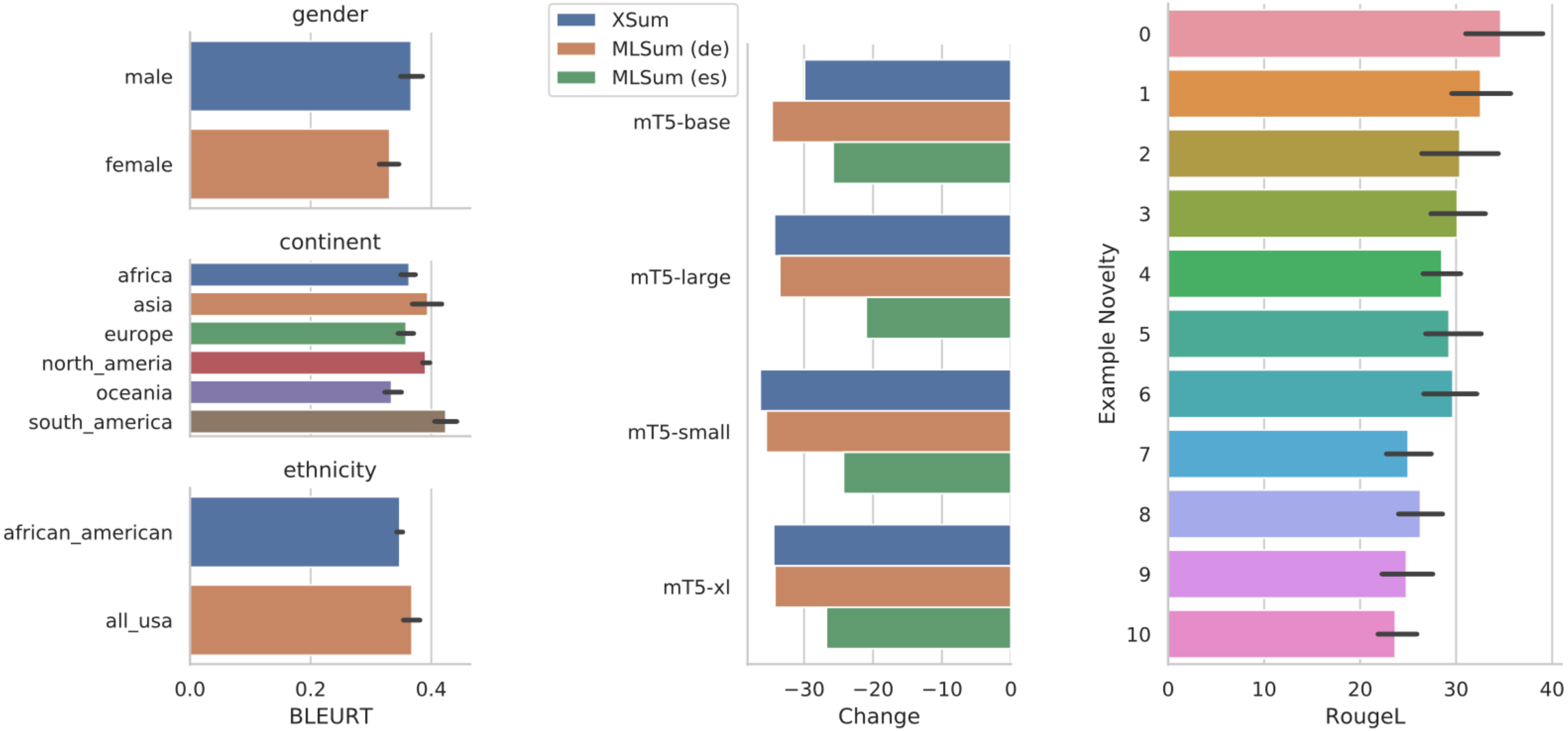


Figure 3: WebNLG results for English and Russian for some subpopulations. The scores of the four models are averaged; error bars indicate variance between model sizes.

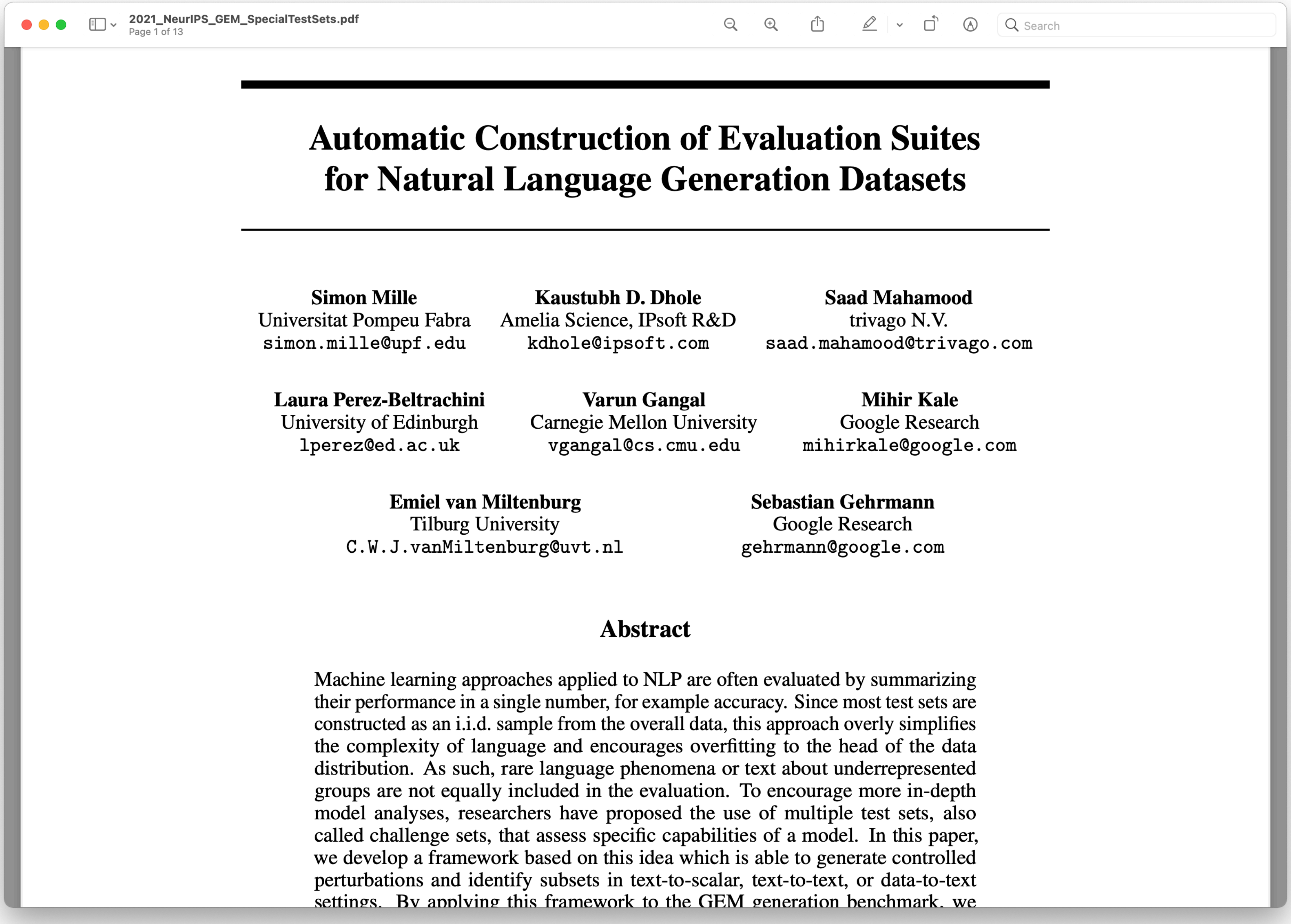


(a) BLEURT results on the fairness-related ToTTo subpopulations.

(b) Relative BERTSCORE change between original and COVID sets.

(c) ROUGE-L score based on fraction of novel words in XSum

Figure 4: The three figures demonstrate the expressiveness of results from challenge sets. In (a), we observe performance differences between subpopulations that talk about people, and in (b) and (c) we demonstrate that models perform poorly when they encounter new concepts and words.



Automatic Construction of Evaluation Suites for Natural Language Generation Datasets

Simon Mille
Universitat Pompeu Fabra
simon.mille@upf.edu

Kaustubh D. Dhole
Amelia Science, IPsoft R&D
kdhole@ipsoft.com

Saad Mahamood
trivago N.V.
saad.mahamood@trivago.com

Laura Perez-Beltrachini
University of Edinburgh
lperez@ed.ac.uk

Varun Gangal
Carnegie Mellon University
vgangal@cs.cmu.edu

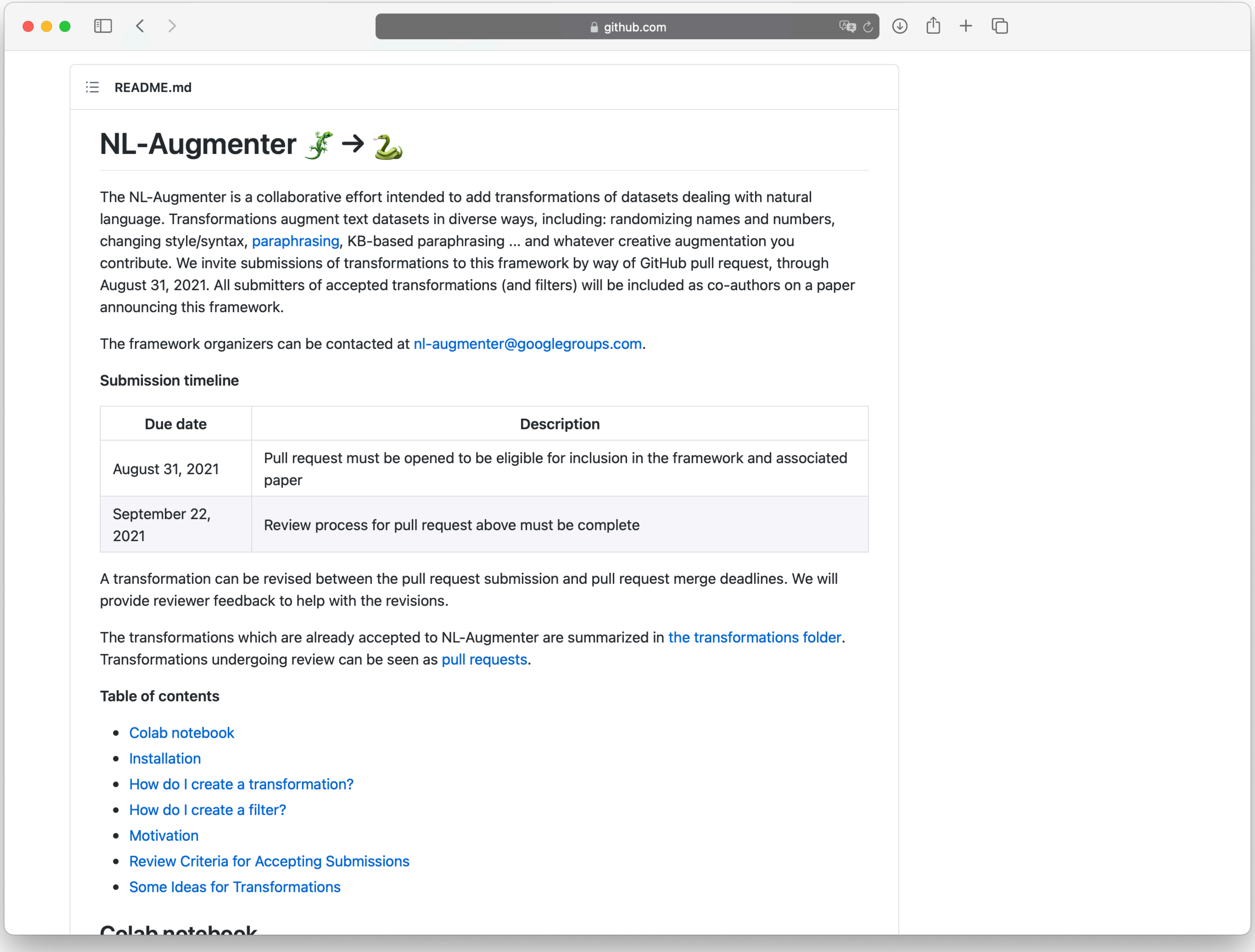
Mihir Kale
Google Research
mihirkale@google.com

Emiel van Miltenburg
Tilburg University
C.W.J.vanMiltenburg@uvt.nl

Sebastian Gehrmann
Google Research
gehrmann@google.com

Abstract

Machine learning approaches applied to NLP are often evaluated by summarizing their performance in a single number, for example accuracy. Since most test sets are constructed as an i.i.d. sample from the overall data, this approach overly simplifies the complexity of language and encourages overfitting to the head of the data distribution. As such, rare language phenomena or text about underrepresented groups are not equally included in the evaluation. To encourage more in-depth model analyses, researchers have proposed the use of multiple test sets, also called challenge sets, that assess specific capabilities of a model. In this paper, we develop a framework based on this idea which is able to generate controlled perturbations and identify subsets in text-to-scalar, text-to-text, or data-to-text settings. By applying this framework to the GEM generation benchmark, we



NL-Augmenter 🦎 → 🐍

The NL-Augmenter is a collaborative effort intended to add transformations of datasets dealing with natural language. Transformations augment text datasets in diverse ways, including: randomizing names and numbers, changing style/syntax, paraphrasing, KB-based paraphrasing ... and whatever creative augmentation you contribute. We invite submissions of transformations to this framework by way of GitHub pull request, through August 31, 2021. All submitters of accepted transformations (and filters) will be included as co-authors on a paper announcing this framework.

The framework organizers can be contacted at nl-augmenter@googlegroups.com.

Submission timeline

Due date	Description
August 31, 2021	Pull request must be opened to be eligible for inclusion in the framework and associated paper
September 22, 2021	Review process for pull request above must be complete

A transformation can be revised between the pull request submission and pull request merge deadlines. We will provide reviewer feedback to help with the revisions.

The transformations which are already accepted to NL-Augmenter are summarized in the transformations folder. Transformations undergoing review can be seen as pull requests.

Table of contents

Colab notebook

4. Human Evaluation Efforts

Human Evaluation Efforts

Efforts to improve Human Evaluation in NLG:

- There is a greater awareness of the issues of human evaluation in NLG.
- Howcroft et al. makes several high-level minimum recommendations when reporting human evaluations in NLG.

SYSTEM	
task	**What problem are you solving (e.g. data-to-text)?** How does it relate to other NLG (sub)tasks?
input/output	**What do you feed in and get out of your system?** Show examples of inputs and outputs of your system. Additionally, if you include pre and post-processing steps in your pipeline, clarify whether your input is to the preprocessing, and your output is from the post-processing, step, or what you consider to be the 'core' NLG system. In general, make it easy for readers to determine what form the data is in as it flows through your system.
EVALUATION CRITERIA	
name	**What is the name for the quality criterion you are measuring (e.g. grammaticality**)?
definition	**How do you define that quality criterion?** Provide a definition for your criterion. It is okay to cite another paper for the definition; however, it should be easy for your readers to figure out what aspects of the text you wanted to evaluate.
OPERATIONALISATION	
instrument type	**How are you collecting responses?** Direct ratings, post-edits, surveys, observation? Rankings or rating scales with numbers or verbal descriptors? Provide the full prompt or question with the set of possible response values where applicable, e.g. when using Likert scales.
instructions, prompts, and questions	**What are your participants responding to?** Following instructions, answering a question, agreeing with a statement? *The exact text you give your participants is important for anyone trying to replicate your experiments.* In addition to the immediate task instructions, question or prompt, provide the full set of instructions as part of your experimental design materials in an appendix.

Table 7: Reporting of human evaluations in NLG: Recommended minimum information to include.

Human Evaluation Efforts

Efforts to improve Human Evaluation in NLG:

- There is a greater awareness of the issues of human evaluation in NLG.
- Howcroft et al. makes several high-level minimum recommendations when reporting human evaluations in NLG.
- To improve reproducibility testing and meta-evaluations Belz et al. introduced a classification system.

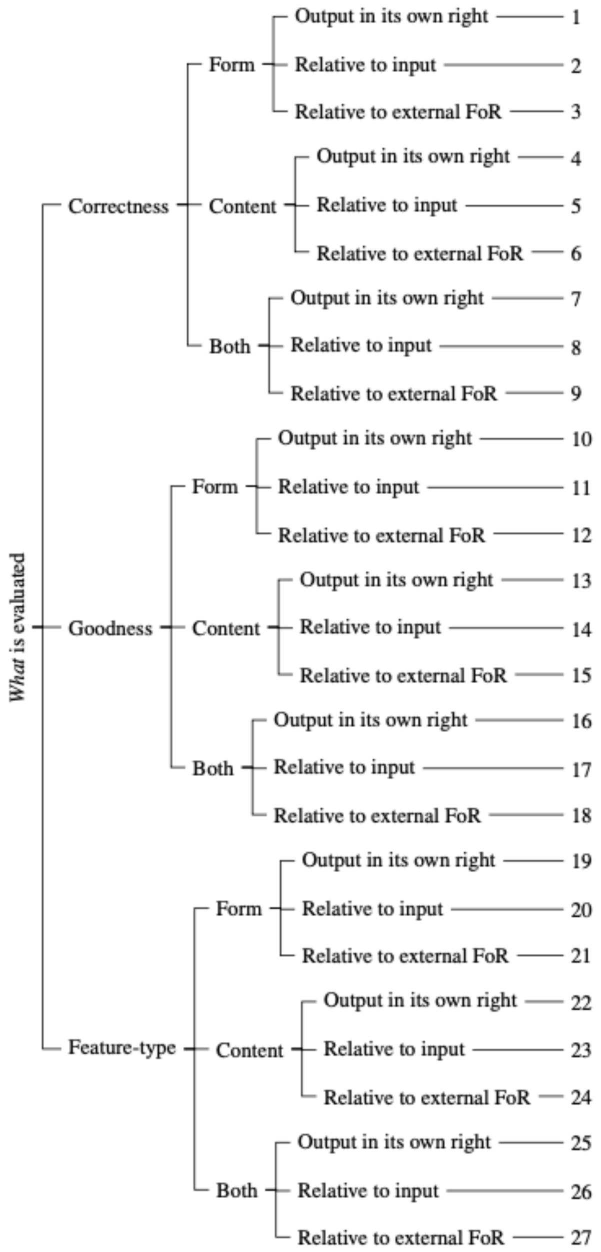


Figure 1: Quality-criterion properties and the 27 different groupings they define (FoR = frame of reference).

Human Evaluation Efforts

Efforts to improve Human Evaluation in NLG:

- There is a greater awareness of the issues of human evaluation in NLG.
- Howcroft et al. makes several high-level minimum recommendations when reporting human evaluations in NLG.
- To improve reproducibility testing and meta-evaluations Belz et al. introduced a classification system.
- Subfields of human evaluation in NLG have also introduced checklists e.g. *Commonsense Evaluation Card (CEC)* for Commonsense NLG human evaluations [Clinciu et al., 2021].
- Still a lot more work to do…

5. Reproducibility Efforts

Reproducibility Efforts

Efforts to improve Reproducibility in NLG:

- ReproGen 2021 Shared Task was the first shared task in NLG to attempt to reproduce results from past NLG human evaluations with four team submissions.
- Results from the shared task showed either small or large percentage differences in reproduced scores depending on the paper being reproduced [Belz et al., 2021].
    - Differences in reproduction cohort could be a contributing factor.
    - Need more information about evaluators and other aspects to conduct reproduction studies.
- Further work ongoing with the 2022 ReproGen Shared Task and ReproHum project.

6. Conclusions

Conclusions

- Automatic and Human evaluations have multiple shortcomings at present within NLG.
- Neural NLG approaches have additional challenges such as factual accuracy, ethics, etc.
- The GEM project aims to address some of these by creating a living benchmark to uncover these shortcomings of neural NLG models.
- By generating challenge sets we can observe the robustness of a given model to perturbations. And allow us to have a better understanding of the shortcomings of a given model.
- Work is underway to improve human evaluation practices in NLG and reproducibility.

Conclusions

More generalisable conclusions…

- There is no perfect one way to conduct an evaluation.
- The quoted performance of a model in a single number or evaluation may not necessarily be the full story.
- All AI models encode biases explicitly and/or implicitly. Therefore it is important to have evaluations on multiple dimensions.
- It is important to probe a given model to appreciate its abilities and limitations with both automatic and human evaluations methods.