A General Benchmarking Framework for Text Generation

Diego Moussallem[1] **Paramjot Kaur**[1] **Thiago Ferreira**[2] **Chris van der Lee**[3]
Anastasia Shimorina[4] **Felix Conrads**[1] **Michael Röder**[1,5] **René Speck**[1] **Claire Gardent**[6]
Simon Mille[7] **Nikolai Ilinykh**[8] **Axel-Cyrille Ngonga Ngomo**[1,5]
[1] Data Science Group, Paderborn University, Germany
[2] Linguistics Department, Federal University of Minas Gerais, Brazil
[3] Tilburg center for Cognition and Communication (TiCC), Tilburg University, The Netherlands
[4] Université de Lorraine / LORIA, France
[5] Institute for Applied Informatics, Germany
[6] CNRS / LORIA, France
[7] Universitat Pompeu Fabra, Spain
[8] Gothenburg University, Sweden

Abstract

The RDF-to-text task has recently gained substantial attention due to the continuous growth of RDF knowledge graphs in number and size. Recent studies have focused on systematically comparing RDF-to-text approaches on benchmarking datasets such as WebNLG. Although some evaluation tools have already been proposed for text generation, none of the existing solutions abides by the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles and involves RDF data for the knowledge extraction task. In this paper, we present BENG, a FAIR benchmarking platform for Natural Language Generation (NLG) and Knowledge Extraction systems with focus on RDF data. BENG builds upon the successful benchmarking platform GERBIL, is open-source and is publicly available along with the data it contains.

1 Introduction

NLG is the process of generating coherent natural language text from non-linguistic data (Reiter and Dale, 2000). A large number of approaches with distinct inputs have been employed for NLG systems over the last years (Gatt and Krahmer, 2018). After having been addressed in only a few papers at the beginning of the last decade (Ell et al., 2012; Ngonga Ngomo et al., 2013), the generation of natural language from Resource Description Framework (RDF) data has gained substantial attention (Gardent et al., 2017b). The RDF-to-text task has hence been proposed to investigate the quality of automatically generated texts from RDF Knowledge Graphs (KGs) (Colin et al., 2016; Moussallem et al., 2020). Recent studies have focused on comparing systematically neural pipeline and end-to-end data-to-text approaches for the generation of text from RDF triples (Ferreira et al., 2019). However, a transparent comparison of RDF-based NLG systems is costly and prone to failure when relying only on benchmarking datasets such as WebNLG without a proper benchmarking platform.

Recent works have hence proposed evaluation tools for text generation such as VizSeq (Wang et al., 2019) and MT-ComparEval (Klejch et al., 2015). However, none of these tools abides by the FAIR principles (Wilkinson et al., 2016), which are now widely regarded as a key first step to ensure reproducible research in scientific experiments. Moreover, none of the tools aforementioned involves RDF data for the knowledge (relations and entities) extraction task (KE).

In this paper, we address this gap by presenting BENG, a FAIR benchmarking platform for NLG and Knowledge Extraction systems. BENG is available as an online instance with a user-friendly interface that can be freely used by researchers to benchmark their systems without the need to set up the benchmarking platform by themselves. Moreover, BENG is an open-source project which can be extended (w.r.t. metrics and systems) and executed locally.[1,2]

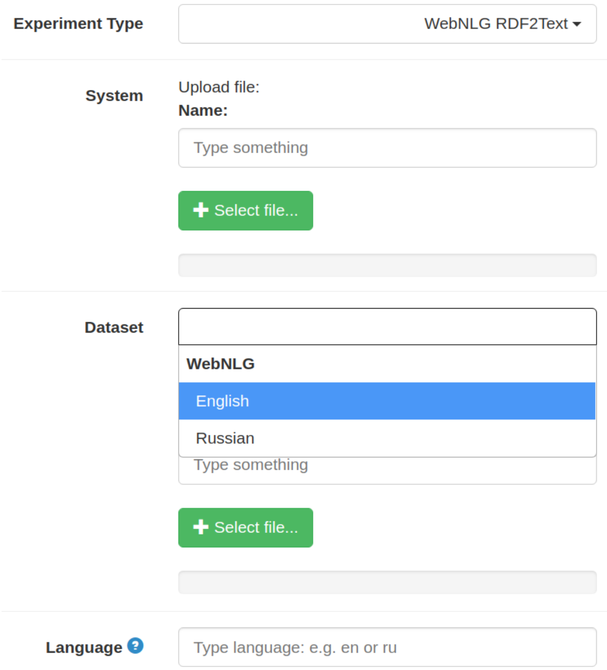
Our benchmarking results show that BENG can foster the development of NLG approaches by providing an easy way to access, compare, and reuse results among NLG systems. Moreover, BENG can support the investigation of multilingual approaches as it contains variations of the WebNLG datasets in languages other than English (German and Russian).

2 Related work

A significant body of research has been devoted to providing evaluation tools for text generation approaches. Recently, Wang et al. (2019) proposed VizSeq as a visual analysis toolkit for independent

[1] https://beng.dice-research.org/

[2] https://github.com/dice-group/BENG



Experiment Type: WebNLG RDF2Text

Figure 1: Experiment Configuration

text generation tasks, for example, Machine Translation (MT) or NLG. VizSeq supports multimodal (images, videos, texts) sources and multiple text references to provide a detailed picture of system evaluations. In MT, compare-mt (Neubig et al., 2019) and MT-ComparEval (Klejch et al., 2015) are related tools for comparative analysis with automatic measures that provide a high-level view of major differences between MT outputs. In turn, Vis-Eval Metric Viewer (Steele and Specia, 2018) and iBLEU (Madnani, 2011) present metric scores as a visual interface. Other tools focus on the interpretability of the text generation process and language model parameters such as the OpenNMT visualisation tool (Klein et al., 2018), LM (Rong and Adar, 2016), and Seq2Seq (Strobelt et al., 2019). Although MT and NLG tasks rely on the same metrics for evaluating their outputs, none of the aforementioned tools rely on FAIR principles for the sake of reproducible research. Therefore, BENG is the first evaluation tool that abides by the FAIR principles for the text generation task.

3 Framework

BENG addresses the problem of comparing different NLG systems using automatic metric results while relying on FAIR principles. It is based on a service-oriented architecture that reuses components from the FAIR benchmarking platform GERBIL (Röder et al., 2018), a benchmarking platform for Named Entity Recognition and Entity Linking systems. We chose GERBIL because it has already been used successfully in more than 80,000 experiments. We reuse the mechanisms implemented by GERBIL to handle large experiments, generate experiment Uniform Resource Identifiers (URIs) and store the results of experiments in a database. Some of the main components of GERBIL (e.g., annotation systems, datasets, matching process, and performance metrics) are replaced by components which abide by the requirements of the evaluation of NLG systems. BENG follows the FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016) as detailed in Table 1.

To cater for the needs of NLG experiments, BENG provides different experiment configurations, which are explained in the following along with their parameters.

3.1 Experiment types

Figure 1 shows the experiment configuration interface which allows users to select the files and evaluate their systems. It has four experiment types:

RDF2Text with two variants: a) WebNLG RDF2Text where the user evaluates the hypothesis on the WebNLG development and test sets. Here, the user just needs to upload the text file in the interface. The evaluation results are generated automatically and represented by a URI. Additionally, all systems which perform this experiment can be found in the leaderboard in Figure 2. b) NLG, where the users have the freedom to upload their own hypothesis and reference files. In this case, results are only displayed under the generated URI, not in the leaderboard as there is no common dataset. Both variants use a python script to evaluate the results with automatic metrics which are discussed in Section 3.2.

Text2RDF with two variants: WebNLG Text2RDF and KE which allow users to evaluate the relation extraction models which convert text into a set of RDF triples. The evaluation script uses Precision, Recall, F1-score as metrics. The evaluation algorithm relies on four types of matches (exact, partial, strict, and Enttype[3]) to compare the candidate triples with the reference triples. In the WebNLG Text2RDF experiment type, the users can upload the candidate triples and select the

[3] Described in details here: https://webnlg-challenge.loria.fr/challenge_2020/#automatic-evaluation

Table 1: A shortened description of the FAIR principles and how BENG addresses each of them.

Findable	F1.	Usage of globally unique, persistent identifiers	Unique W3ID URIs per experiment
	F2.	Data have rich metadata	Experimental configuration as RDF
	F3.	Metadata include the identifier of the data it describes	Relates via RDF
	F4.	(Meta)data are registered or indexed in a searchable resource	Batch-updated SPARQL endpoint
Accessible	A1.	(Meta)data are retrievable using a standardized communications protocol	HTTP (with JSON-LD)
	A1.1	The protocol is open, free, and universally implementable	HTTP is an open standard
	A1.2	The protocol allows for authentication/authorization	Not necessary for BENG
	A2.	Metadata are accessible, even when the data are not available anymore	Each experiment is archived
Interop.	I1.	(Meta)data use a formal, accessible, shared, and broadly applicable language	RDF, DataID, DataCube
	I2.	(Meta)data use vocabularies that follow FAIR principles	Community-based, open vocabularies
	I3.	(Meta)data include qualified references to other (meta)data	Datasets are described using DataID
Reusable	R1.	(Meta)data are richly described	Metrics are relevant to the community
	R1.1.	(Meta)data have a clear licensing	BENG uses LGPL-3.0
	R1.2.	(Meta)data are associated with detailed provenance	Provenance is added to each machine-readable experiment data
	R1.3.	(Meta)data meet domain-relevant community standards	BENG covers a superset of domain-relevant data

Table 2: Datasets. T/S – maximum number of triples per set; D – Domains; EN – English; DE – German; RU - Russian.

Name	**Experiment Type**	**Lang.**	**Texts**	**Sets**	T/S	D
WebNLG2017	RDF2Text	EN	25,298	9,674	7	15
		DE	20,370	7,812	7	15
WebNLG2019	RDF2Text	RU	20,800	5,185	7	9
WebNLG2020	RDF2Text/Text2RDF	EN	45,032	16,677	7	19
		RU	20,800	5,185	7	9

WebNLG reference dataset. The results can be found in the leaderboard (Figure 3). On the other side, the KE experiment type allows the users to upload the candidate triples and reference triples. The results are presented in the generated URI, not in the leaderboard as there is no common dataset.

3.2 Automatic Evaluation Metrics

3.2.1 Metrics for Text Generation

BENG includes the most used metrics according to Gatt and Krahmer (2018) and the metrics which correlate better with human evaluations based on recent findings (Sellam et al., 2020). We briefly explain below the automatic evaluation metrics present in BENG.

3.2.1.1 N-gram-based metrics

BLEU (Papineni et al., 2002) is widely chosen for evaluating text generation outputs due to its low costs. BLEU uses a modified precision metric for comparing the hypotheses with the references. For the sake of comparison, BENG uses two implementations of BLEU: (1) Multi-bleu-detok from Moses,[4] (2) BLEU-NLTK from the NLTK library.[5]

METEOR (Banerjee and Lavie, 2005) relies on semantic features to improve correlation quality between system hypotheses and human references. To this end, METEOR considers the synonymy overlap through a shared WordNet synset of the words to overcome some weaknesses of BLEU and NIST (Doddington, 2002). BENG relies on the latest METEOR version.[6]

chrF++ (Popović, 2015, 2016) proposes the use of character n-gram precision and recall (F-score) for automatic evaluation of text generated outputs. ChrF++ has shown a good correlation with human rankings of different MT outputs, especially for morphologically rich target languages. Additionally, it is language- and tokenisation- independent.[7]

TER (Snover et al., 2006) is different from the aforementioned metrics. TER measures the number of necessary edits in an MT/NLG output to match the reference text exactly. The edits consist of insertions, deletions, substitutions and shift of words, as well as capitalisation and punctuation. The TER score is calculated by computing the number of edits divided by the average referenced words.[8]

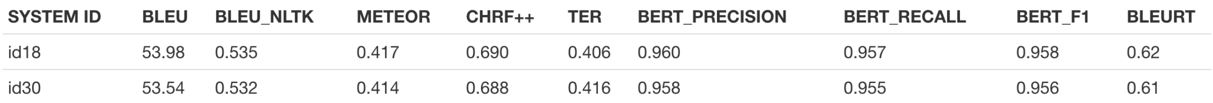
[4] rb.gy/zaffdt

[5] https://www.nltk.org/

[6] rb.gy/6q5zsv

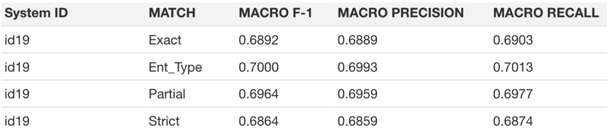
[7] https://github.com/m-popovic/chrF

[8] https://github.com/roy-ht/pyter



SYSTEM ID	BLEU	BLEU_NLTK	METEOR	CHRF++	TER	BERT_PRECISION	BERT_RECALL	BERT_F1	BLEURT
id18	53.98	0.535	0.417	0.690	0.406	0.960	0.957	0.958	0.62
id30	53.54	0.532	0.414	0.688	0.416	0.958	0.955	0.956	0.61

Figure 2: Screenshot of the Leaderboard - RDF2Text



System ID	MATCH	MACRO F-1	MACRO PRECISION	MACRO RECALL
id19	Exact	0.6892	0.6889	0.6903
id19	Ent_Type	0.7000	0.6993	0.7013
id19	Partial	0.6964	0.6959	0.6977
id19	Strict	0.6864	0.6859	0.6874

Figure 3: Screenshot of the Leaderboard - Text2RDF

3.2.1.2 Embedding-based metrics

BERTScore (Zhang et al., 2020) is based on pre-trained BERT contextual embeddings (Devlin et al., 2019). It computes the token similarity of candidate and reference sentences as a sum of cosine similarities between their tokens' embeddings. BERTScore has shown a good correlation with human evaluations through stronger system-level and segment-level correlations than previous metrics.[9]

BLEURT (Sellam et al., 2020) is a learned evaluation metric that relies on BERT (Devlin et al., 2019). It is a novel pre-training scheme that generalises the model by using random disturbance of Wikipedia sentences built up with a diverse set of lexical- and semantic-level supervision signals. In contrast to other recent BERT metrics, BLEURT handles data scarcity in low-resource scenarios.[10]

3.2.2 Metrics for Relation Extraction

Standard evaluation measures are typically applied for evaluating relation extraction systems (Martínez-Rodríguez et al., 2020; Speck and Ngonga Ngomo, 2018; Speck et al., 2018). Thus, we employ three commonly used metrics for each system q. We computed Precision, Recall and F1 Score as follows:

$$Precision(q) = \frac{\text{\# correct system annotations for } q}{\text{\# system annotations for } q};$$

$$Recall(q) = \frac{\text{\# correct system annotations for } q}{\text{\# gold standard annotations}};$$

$$F1\ Score(q) = 2 \cdot \frac{precision(q) \cdot recall(q)}{precision(q) + recall(q)}.$$

3.3 Datasets

BENG includes the WebNLG datasets for the RDF2Text task (refer to Table 2). WebNLG2017 is a semantically varied corpus containing diverse attributes, patterns and shapes. Said corpus (Gardent et al., 2017a,b) consists of 9,674 sets of up to 7 RDF triples in 15 domains mapped to 25,298 target texts. The 15 domains are Astronaut, University, Monument, Building, Comics Character, Food, Airport, Sports Team, Written Work, City, Athlete, Artist, Mean of Transportation, Celestial Body, and Politician. Out of these domains, five (Athlete, Artist, MeanOfTransportation, CelestialBody, Politician) are exclusively present in the test set. The WebNLG2020 datasets are an improvement of the English version of WebNLG2017 and WebNLG2019 for the Russian dataset (Castro-Ferreira et al., 2020). For English, the improvement comprises cleaned texts (around 5,600), added missing triple verbalisations to some texts, and information about tree shapes as well as shape types for each entry.

The German WebNLG version (Castro Ferreira et al., 2018) comprises 20,370 texts describing 7,812 sets of up to 7 RDF triples in 15 domains, while the Russian datasets contain 20,800 texts describing 5,185 sets of up to 7 RDF triples in 9 domains. The English and Russian datasets abide by the criteria to gold standards as several native speakers manually assessed them. The German version can be regarded as a silver standard given that it did not go through the same process and contains some known errors from the Neural Machine Translation (NMT) system used for generating the data. With respect to the Text2RDF task, BENG relies on the same WebNLG 2020 datasets, but uses the triples as a reference and the texts as input.

4 Conclusion

In this paper, we introduced BENG, a general benchmark framework for text generation based on GERBIL's service-oriented architecture. We integrated new experiment types, datasets, and measures. The main advantages of BENG are that it follows the FAIR Guiding Principles and provides a web-based frontend that allows for several use cases enabling lay people and expert users to perform informed comparisons of annotation tools. In

[9] https://github.com/Tiiiger/bert_score

[10] https://github.com/google-research/bleurt

future work, we plan to include other popular NLG benchmarks such as E2E and SR (Belz et al., 2011; Novikova et al., 2017; Mille et al., 2018) and extend the experiment types as well as include the web-services for models instead of uploading the hypotheses.

Acknowledgments

Research funded by the German Federal Ministry of Economics and Technology (BMWI) in the project RAKI (no. 01MD19012D). This work has been supported by the German Federal Ministry of Education and Research (BMBF) within the project DAIKIRI under the grant no 01IS19085B. Also, this work has been supported by the German Federal Ministry for Economic Affairs and Energy (BMWi) within the project SPEAKER under the grant no 01MK20011U. We also would like to thank the funding provided by the Coordination for the Improvement of Higher Education Personnel (CAPES) from Brazil under the grant 88887.367980/2019-00. Finally, we gratefully acknowledge the support of the French National Research Agency (Gardent; award ANR-20-CHIA-0003, XNLG “Multi-lingual, Multi-Source Text Generation”) and of the European Commission in the context of its H2020 Program under the grant numbers 870930-RIA, 779962-RIA, 825079-RIA, 786731-RIA at Universitat Pompeu Fabra.

References

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*, volume 29, pages 65–72.

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 Workshop on Natural Language Generation*, pages 217–226, Nancy, France. Association for Computational Linguistics.

Thiago Castro-Ferreira, Claire Gardent, Chris van der Lee, Nikolai Ilinykh, Simon Mille, Diego Moussalem, and Anastasia Shimorina, editors. 2020. *Proceedings of the 3rd WebNLG Workshop on Natural Language Generation from the Semantic Web (WebNLG+ 2020)*. Association for Computational Linguistics, Dublin, Ireland.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176. Association for Computational Linguistics.

Emilie Colin, Claire Gardent, Yassine Mrabet, Shashi Narayan, and Laura Perez-Beltrachini. 2016. The webnlg challenge: Generating text from dbpedia data. In *Proceedings of the 9th INLG conference*, pages 163–167.

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.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT ’02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Basil Ell, Denny Vrandečić, and Elena Simperl. 2012. Spartiqulation: Verbalizing sparql queries. In *Extended Semantic Web Conference*, pages 117–131. Springer.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for nlg micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit.

Ondřej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. Mt-compareval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, 104(1):63–74.

Nitin Madnani. 2011. ibleu: Interactively debugging and scoring statistical machine translation systems. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 213–214. IEEE.

José-Lázaro Martínez-Rodríguez, Aidan Hogan, and Ivan López-Arévalo. 2020. Information extraction meets the semantic web: A survey. *Semantic Web*, 11(2):255–335.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The first multilingual surface realisation shared task (SR'18): Overview and evaluation results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.

Diego Moussallem, Dwaraknath Gnaneshwar, Thiago Castro Ferreira, and Axel-Cyrille Ngonga Ngomo. 2020. Nabu–multilingual graph-based neural rdf verbalizer. In *International Semantic Web Conference*, pages 420–437. Springer.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Axel-Cyrille Ngonga Ngomo, Lorenz Bühmann, Christina Unger, Jens Lehmann, and Daniel Gerber. 2013. Sorry, i don't speak sparql: translating sparql queries into natural language. In *Proceedings of the 22nd international conference on World Wide Web*, pages 977–988.

Jekaterina Novikova, Ondrej Dusek, 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.

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*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 499–504.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. GERBIL - benchmarking named entity recognition and linking consistently. *Semantic Web*, 9(5):605–625.

Xin Rong and Eytan Adar. 2016. Visual tools for debugging neural language models. In *Proceedings of ICML Workshop on Visualization for Deep Learning*.

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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

René Speck and Axel-Cyrille Ngonga Ngomo. 2018. On extracting relations using distributional semantics and a tree generalization. In *Proceedings of The 21th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2018)*.

René Speck, Michael Röder, Felix Conrads, Hyndavi Rebba, Catherine Camilla Romiyo, Gurudevi Salakki, Rutuja Suryawanshi, Danish Ahmed, Nikit Srivastava, Mohit Mahajan, and Axel-Cyrille Ngonga Ngomo. 2018. Open knowledge extraction challenge 2018. In *Semantic Web Evaluation Challenge*, page 39–51. Springer International Publishing.

David Steele and Lucia Specia. 2018. Vis-eval metric viewer: A visualisation tool for inspecting and evaluating metric scores of machine translation output. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 71–75.

Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2019. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. VizSeq: a visual analysis toolkit for text generation tasks. In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, pages 253–258, Hong Kong, China. Association for Computational Linguistics.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.