Improving Robustness and Reproducibility in Validation of Natural Language Generation Systems

Javier González Corbelle

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain

Abstract

Evaluating properly the goodness of Natural Language Generation (NLG) systems remains a hot topic that requires further research. There are two major issues to tackle: (i) the lack of universal metrics; and (ii) the lack of a rigorous evaluation protocol. In this doctoral dissertation, we began with proposing a novel protocol for validation of NLG systems that is to be grounded in psycholinguistics and inspired from Sensory Evaluation Practices and Analysis. Then, we paid attention to a careful experimental design, regarding the number of assessors, samples, stimuli, and so on, which are required to get statistically sounded robust and reproducible evaluations. More precisely, we focused on analyzing the balance between fidelity and naturalness in texts automatically generated by NLG tools for data to text applications. Our preliminary results show the impact of the lack of domain knowledge in the generation of texts containing input-output divergences in the meteorological domain. We are now in the validation stage, considering new methods to aid human evaluations of generated texts from meteorological data, but we also plan to explore other application domains in the near future.

Keywords

Natural Language Generation, Evaluation, Reproducibility

1. Introduction

Natural Language Generation (NLG) is a subfield within the broader field of Natural Language Processing (NLP) that focuses on generating human-like text from data. The evaluation of NLG systems has recently aroused much interest in the research community since it should address several challenging aspects such as readability of the generated texts, adequacy to the user within a particular context but also linguistic quality related issues (e.g., correctness, coherence, understandability), among others. NLG evaluation plays an important role in the design of language generation systems and has a direct impact on the reliability of such tools in real-world applications (e.g., automatic report generation in business or recommendation systems for medical assistance).

NLG evaluations can be classified regarding two dimensions. On the one hand, according to what we are evaluating we can distinguish between extrinsic (evaluate the impact of the system output on the user or given task) and intrinsic evaluation (evaluate the text quality: readability, consistency, correctness, etc.). On the other hand, according to how the evaluation is carried out, whether by humans or by machines, we can differentiate between human and automatic evaluation [1]. The metrics used to automatically evaluate NLG systems are usually based on comparing the generated texts with a set of reference texts and giving a final similarity score. Such metrics are easy to apply, faster and cheaper than human evaluation. However, it has been proved that the correlation of those metrics with human judgments is low and, therefore, human evaluations are preferable [2]. The procedure of human evaluations consists of resorting to several human judges that are asked to assess different characteristics of the generated texts. Human evaluation gives more reliable results than data-driven automatic metrics but requires more resources [3].

Lee et al. [4] point to data-driven metrics as the most popular evaluation method, even though do not correlate with human judgments. To motivate human evaluation, their paper shows little agreement among researchers on how to perform manual evaluation. However, in the community, there are no

Doctoral Symposium on Natural Language Processing, 26 September 2024, Valladolid, Spain.

j.gonzalez.corbelle@usc.es (J. G. Corbelle)

https://citius.gal/team/javier-gonzalez-corbelle/ (J. G. Corbelle)

0000-0002-9457-0279 (J. G. Corbelle)

formal guidelines on how to evaluate an NLG system properly. Moreover, van Miltenburg et al. [5] refer to another major problem in the context of NLG validation: the under-reporting of errors. Most of the evaluations, either manual or automatic, only give a final score and do not report any details about the actual errors that systems make.

Taking as reference the Sensory Analysis field, a well-established scientific discipline with a wide range of applications (e.g., tasting cheeses, oils, wines, creams, etc.), we can see that there are standards for human evaluation developed by the International Organization for Standardization (ISO) [6]. In the NLG research field, it would be appreciated to have an equivalent solid protocol that ensures the comprehensiveness and reliability of human evaluations and do not merely gives a score. The ideal protocol should report what are the strengths and weaknesses of the system under evaluation and handle the subjectivity introduced by the evaluators when judging system-generated texts. So, with the results of the evaluation, it will be possible to extract conclusions about the performance of the system and try to improve it if needed. Having this protocol would improve the credibility, reproducibility, and robustness of human evaluations.

2. Background

2.1. Data to Text

Data to text (D2T) is a classic NLG task that involves the automatic creation of natural language descriptions from structured data. D2T is focused on transforming non-linguistic data (e.g., tabular data, databases, XML, etc.) into human-readable text. One of the most popular books on NLG, centered in the D2T task, was published by Reiter and Dale [7]. But, since the publication of this pioneering book, new methods have been developed in the field of D2T. Traditionally NLG had to address at least two main tasks: content selection (i.e., selecting the appropriate information to include in the final narrative) and surface realization (i.e., communicating the selected information in the right format). Nowadays the deep learning models have replaced the rule-based or template-based NLG systems [8] and the current end-to-end models are capable of addressing the whole generation pipeline at once, generating more complex outputs by learning lexical and syntactic richness from large corpus and datasets.

End-to-end NLG models outperform classic rule or template based models in terms of fluency, naturalness and variability of generated texts. However, some generated texts have lack of coherence, are unstructured, mention information that is not in the input data (hallucination), or simply ignore some relevant data (omission). Several authors proposed methods to tackle hallucination and/or omission issues, reducing the semantic noise in the training data [9, 10] or enhancing the neural models to learn relevant parts of each data instance [11], but it is not yet possible to assume that end-to-end NLG models will not produce divergences with respect to the input data. Improving the reliability of these type of models remains a challenging research topic.

2.2. NLG Evaluation

Being able to assess the weaknesses and strengths of NLG models can allow us to improve them and therefore generate more reliable text outputs. In the current landscape, automatic versus human evaluation methods remains a topic of intense debate when assessing the outputs of NLG systems.

On the one hand, Reiter [12] demonstrated the shortcomings and poor correlation with human judgements of the commonly used n-gram based automatic metrics, such as BLEU [13] or ROUGE [14]. Accordingly, other metrics have emerged, such us embedding-based metrics to measure similarity between reference and candidate texts like BERTScore [15] or pre-trained metrics (i.e., neural models trained to learn how to automatically do an evaluation task) like BLEURT [16] or NUBIA [17]. There are also recent studies regarding the application of ChatGPT for assessing generation tasks [18]. However, despite these efforts to produce new data-driven automatic metrics, which are inspired from the machine learning community, the lack of correlation with human evaluation persists [19]. On the other hand, human evaluations are the most reliable method to assess NLG systems, but despite some recommended

best practices for human evaluation by Lee et al. [4], the NLG research community is still doing efforts to set the basis for reproducible human evaluation [20, 21]. Stating a formal protocol for carrying out NLG human evaluation is not a straightforward task, since there are multiple factors that must be considered: define textual properties to be assessed, evaluation criteria that human evaluators must follow, number of human evaluators, number of items to evaluate, number of questions per item, statistical tests, tools for data analysis, etc.

This thesis dissertation focuses on tackling the issues of evaluating modern NLG systems through the design of a novel validation protocol following good practices with the aim of ensuring the robustness and reproducibility of the evaluations, which is a critical issue in the NLG field [20].

3. Objectives

The main goal of this PhD thesis is to design, implement and test a new protocol for validation of fidelity-naturalness trade-off in D2T NLG systems. Our hypothesis is that given a specific field, we can achieve a semi-automated protocol for validating the quality of a D2T system, and then generalize that protocol to any other field in which any NLG systems are used. To achieve this goal, the following specific objectives are being addressed:

- Systematic literature review and experimental study of the correlation between data-driven and human-centered intrinsic/extrinsic metrics for measuring fidelity and naturalness of texts in the context of NLG validation. As a result, the definition of guidelines for validation of NLG systems in general and D2T systems in particular.
- Implementation and test of the previous guidelines to validate a pre-existing NLG system in a pilot related to meteorological data, paying attention to two types of users: meteorologists (experts) and end-users. Use of a neural generation system to perform the task of weather forecast generation, and evaluate the generated texts, focusing on the content and form of the textual outputs. The collected data will be cured and released as an open resource for the scientific community.
- Design, implement, deploy, and validate a D2T system for the same application problem (reporting weather forecast) previously considered. We will consider the combination of end-to-end deep learning approaches with rule-based NLG pipelines to achieve a good fidelity-naturalness trade-off, and we will conduct a thorough evaluation of the system following the proposed guidelines.

4. Methodology

We are following an iterative research methodology that is grounded on software engineering principles. We apply the principles of agile software development, involving users in the design process, pursuing continuous improvement, encouraging rapid and flexible response to changes, and supporting frequent delivery of functional software along with the related technical and user manuals. For each scientific-technical objective previously stated, our research methodology comprises several sprints with the following steps:

1. **Requirements Specification.** The aim is to carefully define the research objective. It must be clear what are the final goals of our task. E.g., design a D2T automatic weather forecast generator.
2. **Literature Review.** The bibliographic study aimed to identify related work. A careful analysis of the advantages and drawbacks of existing approaches. E.g., analyse in the state of the art which are the main problems to deal with when designing a D2T system (hallucinations, omissions, etc.) and also the best way to evaluate the proposed system.
3. **Implementation.** Design and development of a new proposal aimed at producing significant advances with respect to the state of the art. The implementations must be ready to be validated empirically in the next step. E.g., create a system to generate meteorological texts from structured data, using a state of the art approach to reduce the generation of input-output divergences.

4. **Validation.** This step includes both verification and validation, but also refinement of the implemented proposal if needed in the light of the experimental results. E.g., validation of the texts generated by the model with a tool to detect input-output divergences.
5. **Testing and review.** The new proposal, once validated, will be tested in a real use case. With the results obtained in the testing phase, we can extract conclusions as to which aspects can be improved in future implementations. E.g., verify with experts on the meteorological field the quality of the generated texts and the severity of the input-output divergences detected during the validation step.

5. Research

5.1. Dealing with Hallucinations and Omissions in D2T

To address the thesis objectives previously defined, we started our research through a systematic literature review in order to identify the state of the art of the evaluation in NLG. One of the issues that we found that had not yet been addressed in depth was the detection and evaluation of divergences between the input and output of NLG end-to-end systems (i.e., hallucinations and omissions). This led us to perform our first experiment over a D2T end-to-end system, in order to analyze hallucinations and omissions in a particular use case. In our work [22], we show the impact of the lack of domain knowledge in the generation of texts containing input-output divergences through a use case on meteorology. The objective of a D2T system in this field is to get as input tabular data from a meteorological database and automatically generate a text describing the data, i.e., generate a meteorological forecast. This facilitates the work of meteorologists, who only have to review the texts generated by the system instead of writing them from scratch.

To analyze the input-output divergences generated by a system, we adapted a D2T Transformer-based model optimized to minimize the appearance of hallucinations to our specific domain, i.e., meteorology. We trained it with a new dataset and corpus curated by meteorologists[1]. The dataset we introduced in this work is made up of 3,033 records of meteorological tabular data along with handwritten textual descriptions in Spanish, provided by MeteoGalicia, the Official Meteorological Agency of Galicia[2]. After training the model with the MeteoGalicia dataset using 70% of the data for training, 15% for validation, and 15% for test, we finally obtained some output texts to analyze, focusing on the generated input-output divergences, that was the aim of our research in this work. We obtained a total of 272 generated texts to analyze from the test set.

We first performed a divergences' detection step with a simple detector in order to identify the clearest divergences, especially those involving hallucinations. The divergence detector designed is a software application composed of two independent parts, one for detecting each type of divergence. On the one hand, the omission detection part works as follows: it looks first at the table with input data values and then checks if all these values are mentioned in the generated text. The detector counts as omission each value which is in the input data but is not explicitly referred to in the output text. On the other hand, the hallucination detection part follows the other way round. It looks first to the output text, identifies all data values which are mentioned in the text, and then checks if they are also included in the related input data. The detector counts as hallucination each value that is mentioned in the output text but is not present in the input data. It must be noted that the weather data from the tables are only related to the state of the sky, so the values are categorical (e.g., "rain", "clouds", etc.) and there is a defined list of possible values, which facilitates the detection of hallucinations and omissions for this specific context.

After applying our divergences detector, we found a very high number of omissions, i.e., 160 out of 272 texts (58%). In order to identify which omissions could be admissible for humans and therefore should not be reported as unacceptable by our detector, we asked a meteorologist to analyze in detail a

[1] https://gitlab.citius.gal/gsi-nlg/meteogalicia-es
[2] https://www.meteogalicia.gal

group of randomly selected cases among the detected omissions. He confirmed that many of them were admissible because in the context of meteorology missing some information is not so severe as it may be in other application domains. In fact, in some cases, the meteorologist preferred certain omission to the exhaustive verbalization of all the values in the data table what could lead to a long, verbose, repetitive and less natural text. Regarding the hallucinations, our detector identified a total of 46 hallucinations that were later analyzed by an expert in the meteorology domain, with the aim of classifying them by severity (i.e., admissible, partially admissible, or inadmissible), taking into account the domain knowledge. Surprisingly, the expert rated as admissible 12 out of all the 46 detected hallucinations. Formally speaking, all these cases were hallucinations (i.e., the values mentioned in the output were not present in the input data) but, according to the meteorologist's background and in agreement with contextual information and commonsense reasoning, they were admissible.

After this experimentation we were able to draw some important remarks. Neural D2T systems, after being trained with large-scale datasets, can generate natural and fluid texts, but more often than not the generated texts provide unfaithful information or inconsistencies with respect to the input data, mainly in the form of omissions and/or hallucinations. We must take into account that in practice, meteorologists rely on contextual information and commonsense reasoning beyond input data when writing weather forecasts. Current neural D2T systems can not capture such a general knowledge because they are only guided by the given training data, which is a limitation of this type of models and/or learning methodology. This means that for truly complex tasks, where either omissions or hallucinations may be critical, neural models have to be endowed and integrated with other knowledge sources different from data, if we want them to achieve high quality automatically generated texts which are as correct as expert-made ones. Last but not least, the high level of naturalness and fluidity that neural D2T systems usually achieve may raise too high expectations in end users, who may be frustrated when discovering some misleading pieces of information. As future work, we plan in the midterm to enrich our neural D2T system with a knowledge base including meteorological facts (regarding both spatial and temporal references) but also in the longterm with temporal knowledge. As a result, we expect to improve both text generation and hallucination/omission detection.

5.2. Impact of the Number of Items in NLG Human Evaluation

Human evaluation of neural models in NLG requires a careful experimental design in terms of the number of evaluators, number of items to assess, number of quality criteria, among other factors, for the sake of reproducibility as well as for ensuring that significant conclusions are drawn. One of the objectives of this thesis is to design a set of guidelines for validation of NLG systems, so studying the impact of choosing different parameters when performing a human evaluation is crucial. Hence, in our work [23] we addressed empirically the impact of the number of items to assess in the context of human evaluation of NLG systems.

We wanted to test the following hypothesis: "well-known resampling statistical methods can contribute to getting significant results even with a small number of items to be evaluated by each evaluator". To do so, we first carried out an evaluation with three different raters on a pool of texts generated by a D2T neural system. Three evaluators rated a total of 246 texts each in a 5-point Likert scale, based on their fluency, naturalness, and content. Then, with the scores obtained from the evaluation of all the texts, we created different prototypical evaluator profiles grounded on the previous human evaluations. We transformed the three evaluator responses into three categories (i.e., negative, fair and positive) and aggregated them into a global curated score. Finally, from that global score we re-defined five 5-value distributions following the evaluation tendencies characteristic of each of the following prototypical profiles: polarized, neutral, negative, positive, and random profile. Therefore, we had three distributions of scores from real evaluators and five from the prototypical profiles.

Using resampling methods (i.e., bootstrap and resampling without replacement) over all the score distributions, we simulated evaluations in the search for the minimal number of items that is required to get sounded insights and test our hypothesis. We compared the score distributions obtained by evaluating only a limited set of items with the distribution obtained by evaluating all outputs of the

system for a given test set. This way, if we obtain the same score distribution with a smaller test set, we can confirm our hypothesis. We did this for all the prototypical profiles and evaluators.

After carrying out the experimentation and analyzing the results, we concluded that in our use case it is possible to approximate the distribution of evaluations of a real set of texts from a smaller subset of evaluated items. To be more specific, in our experiment with a test set of 246 items and evaluating each text using a 5-point Likert scale, it would be sufficient to evaluate 24 items instead of the whole test set to ensure that, no matter the evaluator profile, we obtain a score distribution equivalent to evaluating the whole set of texts in at least 90% of the cases. This empirical findings validated our initial research hypothesis: "well-known resampling statistical methods can contribute to getting significant results even with a small number of items to be evaluated by each evaluator". As future work, we plan to extend the empirical study to other types of evaluations in which the scoring criteria and scale may vary from those tested in this work. Moreover, we will consider how resampling methods can be integrated in the evaluation procedure to address the lack of resources (e.g., evaluators availability) in NLG human evaluation.

5.3. Reproducibility of Human Evaluations in NLP

In line with the objectives described for the thesis, we also participated in two reproducibility experiments within the REPROHUM project [3]. The key goals of this project are the development of a methodological framework for testing the reproducibility of human evaluations in NLP, and of a multi-lab paradigm for carrying out such tests in practice, being the first study of this kind in NLP. Our participation in the project consisted of two experiments in which we had to reproduce the human evaluations performed in the validation of two different NLG systems, following a protocol including a template for recording the details of individual human evaluation experiments in NLP, i.e., the Human Evaluation Datasheet (HEDS) [24].

The first reproduction experiment was an evaluation of a D2T model that generates sport summaries from NBA basketball game statistics [25]. In the evaluation of this system, the presented model is compared with other four baseline systems. In each of the 300 tasks performed by Amazon Mechanical Turk workers during the evaluation process, they were shown the NBA statistics data table along with a text generated by different D2T models describing the statistics. Workers were asked to count the supported and contradicting facts in the text. Therefore, at the end of the evaluation, there was a score table with the average number of supported and contradicting facts for the five different models tested (i.e., the presented system and the four baselines). Results of this reproduction experiment, showed that the tendency of the scores regarding the supported facts was similar to the original evaluation, while for the contradicted facts all the systems achieved higher scores in our reproduction experiment than in the original one.

The second experiment within the REPROHUM project aimed at reproducing the human evaluation of a controlled text generation method [26]. The evaluation consisted of a comparison of the continuations generated by the proposed method against four different baseline methods designed for the same task. In the evaluation, workers were shown two continuations of a given prompt and they had to choose which one they prefer in terms of fluency, topicality and toxicity, having also the "no preference" option. In each of the 960 evaluation tasks, continuations of different methods were confronted, so that the final results show a comparison of the evaluated method against each of the four baseline methods. After reproducing this evaluation, we compared the original versus our reproduction results and found several differences. In general, in the reproduction experiment the use of the "no preference" option decreased, showing more polarized results. Furthermore, the highest coefficients of variation with respect to the original evaluation were reported for the fluency and toxicity criteria.

In light of the findings from these two reproduction experiments, we can conclude that reproducing a human evaluation in NLP is not a trivial task. Aside from whether or not the results of the reproduction experiment were the same as the original, our work highlights the critical importance of providing

[3] https://reprohum.github.io/

comprehensive details about human evaluations in NLP in order to achieve high reproducibility. Defining correctly the criteria to assess, the number of questions, the type of scales, the sample size, or applying statistical analysis on the results are things that must be considered prior to evaluation and should be reported thoroughly. The standardization of reporting practices for human evaluations, through the use of the HEDS, creates a unified approach that enhances the reproducibility, credibility, and reliability of research endeavors. Our participation in the REPROHUM project, allowed us to learn all these lessons and gain experience in the design of human evaluations, an expertise that will undoubtedly be very useful in our future research. Thus, we also encourage researchers to thoroughly document their NLP evaluations using these guidelines, with the objective of augmenting the quality of contributions in the field.

6. Specific Research Elements Proposed for Discussion

This thesis focuses on the development of a protocol for the evaluation of NLG systems, starting from a specific use case on D2T for the field of meteorology. The research conducted so far has focused on analyzing the challenges of evaluating end-to-end D2T systems, the influence of the number of items in human evaluations, and methods to make these evaluations more reproducible. From this, some aspects for discussion are proposed, related to each of the parts of our research presented in section 5.

6.1. Dealing with Hallucinations and Omissions in D2T

- Generalization of the hallucination detector beyond weather forecasting: Is it possible to design an automatic hallucination detector for NLG systems, or at least for D2T systems in general?
- Can we design a pipeline combining rule-based and end-to-end models leveraging the advantages of each approach (i.e., the content reliability of rule-based models and the fluidity, naturalness, and variety of the texts generated by end-to-end models)?

6.2. Impact of the Number of Items in NLG Human Evaluation

- What human evaluation parameters would be interesting to evaluate their impact?
- Can resampling methods be used for addressing the lack of human resources in human evaluations (i.e., evaluate a small sample of items and extract results equivalent to evaluating the whole set of items)?

6.3. Reproduciblity of Human Evaluations in NLP

- What do we mean by quality criterion of a generated text? How can we measure it correctly?
- Are aspects like readability or complexity of a given text measurable in a human evaluation? Can we measure them automatically?

Acknowledgments

This research work is supported under Grant TED2021-130295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR”, but also under Grants PID2020-112623GB-I00 and PID2021-123152OB-C21 funded by MCIN/AEI/10.13039/501100011033 and by “ESF Investing in your future”. We also acknowledge the support of the Galician Ministry of Culture, Education, Professional Training and University (Grants ED431C2022/19 co-funded by the European Regional Development Fund, ERDF/FEDER program, and Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04) and the Nós Project (Spanish Ministry of Economic Affairs and Digital Transformation and the Recovery, Transformation, and Resilience Plan- Funded by the European Union- NextGenerationEU, with reference 2022/TL22/00215336).

References

[1] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, E. Krahmer, Best practices for the human evaluation of automatically generated text, in: K. van Deemter, C. Lin, H. Takamura (Eds.), Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 355–368. URL: https://aclanthology.org/W19-8643. doi:10.18653/v1/W19-8643.

[2] J. Novikova, O. Dušek, A. Cercas Curry, V. Rieser, Why we need new evaluation metrics for NLG, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2241–2252. URL: https://aclanthology.org/D17-1238. doi:10.18653/v1/D17-1238.

[3] E. Reiter, R. Robertson, L. M. Osman, Lessons from a failure: Generating tailored smoking cessation letters, Artificial Intelligence 144 (2003) 41–58. doi:10.1016/s0004-3702(02)00370-3.

[4] C. Lee, A. Gatt, E. van Miltenburg, E. Krahmer, Human evaluation of automatically generated text: Current trends and best practice guidelines, Computer Speech Language 67 (2021) 101151. URL: https://www.sciencedirect.com/science/article/pii/S088523082030084X. doi:https://doi.org/10.1016/j.csl.2020.101151.

[5] E. van Miltenburg, M. Clinciu, O. Dušek, D. Gkatzia, S. Inglis, L. Leppänen, S. Mahamood, E. Manning, S. Schoch, C. Thomson, L. Wen, Underreporting of errors in NLG output, and what to do about it, in: A. Belz, A. Fan, E. Reiter, Y. Sripada (Eds.), Proceedings of the 14th International Conference on Natural Language Generation, Association for Computational Linguistics, Aberdeen, Scotland, UK, 2021, pp. 140–153. URL: https://aclanthology.org/2021.inlg-1.14. doi:10.18653/v1/2021.inlg-1.14.

[6] AENOR, Análisis sensorial, 2 ed., AENOR (Agencia española de Normalización y Certificación), 2010.

[7] E. Reiter, R. Dale, Building applied natural language generation systems, Natural Language Engineering 3 (1997) 57–87. doi:10.1017/S1351324997001502.

[8] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 65–170. doi:10.1613/jair.5714.

[9] O. Dušek, D. M. Howcroft, V. Rieser, Semantic noise matters for neural natural language generation, in: K. van Deemter, C. Lin, H. Takamura (Eds.), Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 421–426. URL: https://aclanthology.org/W19-8652. doi:10.18653/v1/W19-8652.

[10] H. Wang, Revisiting challenges in data-to-text generation with fact grounding, in: K. van Deemter, C. Lin, H. Takamura (Eds.), Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 311–322. URL: https://aclanthology.org/W19-8639. doi:10.18653/v1/W19-8639.

[11] C. Rebuffel, M. Roberti, L. Soulier, G. Scoutheeten, R. Cancelliere, P. Gallinari, Controlling hallucinations at word level in data-to-text generation, 2021. arXiv:2102.02810.

[12] E. Reiter, A structured review of the validity of BLEU, Computational Linguistics 44 (2018) 393–401. URL: https://aclanthology.org/J18-3002. doi:10.1162/coli_a_00322.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[14] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations (ICLR), OpenReview, 2020.

URL: https://openreview.net/forum?id=SkeHuCVFDr.

[16] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704. doi:10.18653/v1/2020.acl-main.704.

[17] H. Kane, M. Y. Kocyigit, A. Abdalla, P. Ajanoh, M. Coulibali, Nubia: Neural based interchangeability assessor for text generation, 2020. arXiv:2004.14667.

[18] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, J. Zhou, Is ChatGPT a good NLG evaluator? a preliminary study, in: Y. Dong, W. Xiao, L. Wang, F. Liu, G. Carenini (Eds.), Proceedings of the 4th New Frontiers in Summarization Workshop, Association for Computational Linguistics, Hybrid, 2023, pp. 1–11. URL: https://aclanthology.org/2023.newsum-1.1.

[19] F. Moramarco, A. Papadopoulos Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, A. Savkov, Human evaluation and correlation with automatic metrics in consultation note generation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5739–5754. URL: https://aclanthology.org/2022.acl-long.394. doi:10.18653/v1/2022.acl-long.394.

[20] A. Belz, C. Thomson, E. Reiter, S. Mille, Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 3676–3687. URL: https://aclanthology.org/2023.findings-acl.226. doi:10.18653/v1/2023.findings-acl.226.

[21] A. Belz, A Metrological Perspective on Reproducibility in NLP, Computational Linguistics 48 (2022) 1125–1135. URL: https://doi.org/10.1162/coli_a_00448. doi:10.1162/coli_a_00448.

[22] J. González Corbelle, A. Bugarín-Diz, J. Alonso-Moral, J. Taboada, Dealing with hallucination and omission in neural natural language generation: A use case on meteorology., in: S. Shaikh, T. Ferreira, A. Stent (Eds.), Proceedings of the 15th International Conference on Natural Language Generation, Association for Computational Linguistics, Waterville, Maine, USA and virtual meeting, 2022, pp. 121–130. URL: https://aclanthology.org/2022.inlg-main.10. doi:10.18653/v1/2022.inlg-main.10.

[23] J. González-Corbelle, J. M. Alonso-Moral, R. M. Crujeiras, A. Bugarín-Diz, An empirical study on the number of items in human evaluation of automatically generated texts, Procesamiento del Lenguaje Natural 72 (2024) 45–55. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6577.

[24] A. Shimorina, A. Belz, The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP, in: A. Belz, M. Popović, E. Reiter, A. Shimorina (Eds.), Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 54–75. URL: https://aclanthology.org/2022.humeval-1.6. doi:10.18653/v1/2022.humeval-1.6.

[25] J. González Corbelle, J. Alonso, A. Bugarín-Diz, Some lessons learned reproducing human evaluation of a data-to-text system, in: A. Belz, M. Popović, E. Reiter, C. Thomson, J. Sedoc (Eds.), Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 49–68. URL: https://aclanthology.org/2023.humeval-1.5.

[26] J. González Corbelle, A. Vivel Couso, J. M. Alonso-Moral, A. Bugarín-Diz, ReproHum #0927-3: Reproducing the human evaluation of the DExperts controlled text generation method, in: S. Balloccu, A. Belz, R. Huidrom, E. Reiter, J. Sedoc, C. Thomson (Eds.), Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 153–162. URL: https://aclanthology.org/2024.humeval-1.15.