SemEval shared task proposal: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes

Elaine Zosa[♠] **Raúl Vázquez**[♠] **Jörg Tiedemann**[♠] **Vincent Segonne**[♢]
Alessandro Raganato[♡] **Timothee Mickus**[♠] **Marianna Apidianaki**[♣]

[♠] University of Helsinki
[♢] Université Grenoble Alpes
[♡] University of Milano-Bicocca
[♣] University of Pennsylvania

{firstname.lastname}@{[♠]helsinki.fi,[♢]univ-grenoble-alpes.fr,[♡]unimib.it}
[♣]marapi@seas.upenn.edu

1 Overview

The modern NLG landscape is plagued by two interlinked problems: On the one hand, our current neural models have a propensity to produce inaccurate but fluent outputs; on the other hand, our metrics are most apt at describing fluency, rather than correctness. This leads neural networks to "hallucinate", i.e., produce fluent but incorrect outputs that we currently struggle to detect automatically. For instance, Dopierre et al. (2021) report that when trying to produce a paraphrase for the input "*I am not sure where my phone is*", they obtain the following 'hallucination' behavior: "*How can I find the location of any Android mobile*". For many NLG applications, the correctness of an output is however mission critical. For instance, producing a plausible-sounding translation that is inconsistent with the source text puts in jeopardy the usefulness of a machine translation pipeline. With our shared task, we hope to foster the growing interest in this topic in the community (e.g., Ji et al., 2023; Raunak et al., 2021; Guerreiro et al., 2022; Xiao and Wang, 2021; Guo et al., 2022).

In particular, with SHROOM we adopt a *post hoc* setting, where models have already been trained and outputs already produced: participants will be asked to perform binary classification to identify cases of **fluent overgeneration hallucinations** in two different setups: **model-aware** and **model-agnostic** tracks. That is, participants must detect grammatically sound outputs which contain incorrect or unsupported semantic information, inconsistent with the source input, with or without having access to the model that produced the output. To that end, we will provide participants with a collection of checkpoints, inputs, references and outputs of systems covering five different NLG tasks: definition modeling (DM, Noraset et al., 2017), machine translation (MT), paraphrase generation (PG), text simplification (TS) and text summarization (Sum) trained with varying degrees of accuracy. The development set will provide binary annotations from at least five different annotators and a majority vote gold label.

SHROOM will attract the interest of diverse NLG research communities, including QA, dialog and MT. It is related to fact-checking but, instead of analysing claims made by human authors, SHROOM focuses on fluent overgeneration of system outputs. Consequently, aspects like verifiability, check-worthiness and misleading statements are less relevant to SHROOM, whereas the naturalness and fluency of the produced output are more prevalent. Other communities that this task could attract include explainable NLP or uncertainty modeling, since participants will need to account for deviant model outputs. The first edition of SHROOM will also pave the way for follow-up shared tasks and evaluation campaigns. We hope to organize future related tasks on token-level over-generation mistakes detection, as well as broaden the scope of languages and NLG tasks considered.

2 Theoretical framing

Guerreiro et al. (2022) propose a taxonomy of hallucinations that includes oscillatory productions, and fluent but strongly or fully "detached" outputs. While this taxonomy is well constructed, we find it inadequate for the needs of the community at large for four reasons: *(i)* It conflates some issues of fluency with semantic correctness (oscillatory productions are cases of non-fluent overgeneration where no extraneous semantic material is introduced); *(ii)* It only considers the most extreme cases of hallucinations (strongly or fully detached productions), whereas diagnosis of intermediary cases is bound to be more challenging and useful to the community; *(iii)* It focuses only on MT, although other tasks are also known to suffer from fluent overgeneration (e.g., Rohrbach et al., 2018), including the ones we propose to address; *(iv)* It

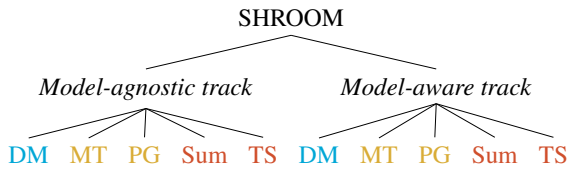


Figure 1: Shared task overview. Datapoints from systems in blue correspond to target-referential datapoints, in red, source-referential, in yellow, either.

uses only lowest scoring outputs, whereas any tool built to verify system outputs ought not to flag non-pathological outputs.

We therefore focus on cases of fluent overgeneration. Judgments pertaining to the over-generative nature of a production can be elicited by means of **inferential semantics**: if an output cannot be inferred from its semantic reference, then it contains some information that is not present in the reference—i.e., the model has generated more than we expected. We will provide multiple annotations and a gold majority label, given the low consensus on semantic annotations (Nie et al., 2020).

An overview of the task is provided in Fig. 1. SHROOM is framed around two key distinctions: (i) model-aware vs. model-agnostic approaches, and (ii) source-referential vs. target-referential datapoints. The former corresponds to whether participants have access to the model that generated the item: **Model-agnostic** approaches are practical, as models may not be accessible to end users; **Model-aware** approaches can lead to richer and more accurate diagnoses. The latter is a consequence of our inferential take on over-generation: what can effectively serve as a semantic reference varies across NLP systems. For **source-referential** datapoints such as those produced by Sum. models, the target is expected to be semantically implied by the source—whereas the converse is not true. For **target-referential** datapoints (e.g. DM, where we fine-tune a language model to produce a definition for a given example of usage) the target is the sole usable semantic reference. In tasks such as PG or MT, where source and target are equivalent, this distinction bears no weight.

An example datapoint displaying how we plan to encode all relevant information in a JSON format is provided in Fig. 2. The datapoint keeps track of the source provided to the model as input (`src`), the intended target (`tgt`), the model production (`hyp`), the type of model this production was derived from (`modeltype`), which can correspond to DM, MT, PG, Sum or TS), whether this datapoint is source- or target-referential (`reference`), as well as the annotations and the gold label (`annotations` and `label`). In the model-aware track, we will also provide a HuggingFace model name (`ckpt`).

```
{
    "src": "It has also been found very useful in
        certain industries that require large amounts
        of <def>process hot water</def>, hence the
        interest of Mohawk Paper Mills.",
    "tgt": "Hot water for use in industrial
        processes.",
    "hyp": "Hot water used in industry."
    "modeltype": "DM",
    "reference": "tgt",
    "id": 42,
    "annotations": [0, 0, 0, 1, 1],
    "label": 0,
    "ckpt": "HelsinkiNLP/DM_step_10K"
}
```

Figure 2: Example target-referential datapoint for the model-aware track.

The test sets for the model-aware and the model-agnostic tracks will have partial overlap, so as to allow us to compare submissions on both tracks after the competition. Submissions will be evaluated according to two criteria: the **accuracy** that the system reached on the binary classification, and the Spearman **correlation** of the systems' output probabilities with the proportion of the annotators marking the item as overgenerating.

3 Data

All SHROOM data (models, outputs and annotations) will be available under a CC-BY license.

Data & model provenance Participants will be provided with generated outputs from multiple systems trained to generate English output at various stages of their training, stemming from five sequence-to-sequence NLG tasks: DM, MT, PG, TS and Sum. We already generated outputs for DM with the architecture of Bevilacqua et al. (2020) over DBnary (Sérasset, 2015), and for MT, using `marian` (Junczys-Dowmunt et al., 2018) on the Tatoeba corpus (Tiedemann, 2020).

Annotation We plan to annotate at least 4,000 items, which will then be split 25%–75% between development and test sets. We will provide additional model outputs to participants as supplementary unannotated data as an unlabeled training split. Participants will have access to a large portion of outputs from the NLG systems and the full set of possible target references to allow corpus-wide approaches. We plan to preselect fluent items so as

	MT		Summarization		DM	
	Src. ref.	Tgt. ref.	Src. ref.	Tgt. ref.	Tgt. ref.	Avg
COMET-QE	78.85 / 81.81	77.71 / 81.06	55.00 / 55.31	58.00 / 56.38	71.50 / 69.80	68.09 / 68.87
COMET	**85.14 / 89.39**	**83.42 / 87.12**	54.00 / 54.25	**63.00 / 61.70**	**76.87 / 75.87**	**72.48 / 73.66**
Seq-Logprob	73.71 / 78.03	69.71 / 74.24	60.00 / 60.63	58.00 / 56.38	73.87 / 72.83	67.05 / 68.42
Attn-ign-SRC	64.57 / 67.42	57.14 / 60.60	**61.00 / 61.70**	58.00 / 57.44	71.12 / 68.74	62.36 / 63.18
MC-DSim	73.14 / 78.03	72.00 / 74.24	56.00 / 55.31	58.00 / 57.44	71.12 / 68.74	66.05 / 66.75
Majority class	61.71 / 66.66	54.28 / 59.48	54.00 / 54.25	58.00 / 56.38	71.12 / 68.74	59.82 / 61.02

Table 1: Optimal accuracy of baseline methods on all / fluent outputs.

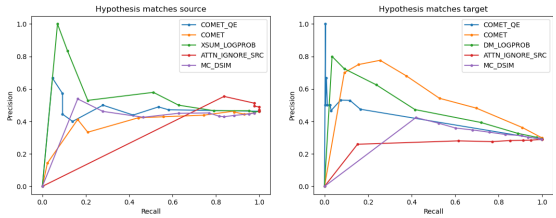
to guarantee a gradient in quality as measured by automated metrics. Items will be annotated by five annotators on whether the reference entails the output. Annotations will be binary, for ease of dataset construction. Gold labels will be defined with respect to the annotators' majority vote.

4 Pilot study

We showcase the relevance of this shared task through a pilot study, where we apply methods proposed by Guerreiro et al. (2022). Model outputs were selected from three NLG tasks: (1) 175 examples for FR-EN MT using a MarianMT model[1]; (2) 100 examples for Sum using Pegasus (Zhang et al., 2019)[2]; (3) 800 examples for DM using Bevilacqua et al. (2020)'s system. We annotated (a) whether the hypothesis is supported by the source; (b) whether the hypothesis matches the target; (c) whether the hypothesis is fluent.

We apply hallucination detection heuristics from Guerreiro et al. on the pilot data. We use COMET and COMET-QE (Rei et al., 2020a,b) as model-agnostic methods. For model-aware methods, we consider the sequence log-probabilities of the hypothesis (Seq-Logprob), the proportion of source tokens that receive low attention mass (Attn-ign-SRC), and the average similarity of the original hypothesis to new hypotheses generated with Monte Carlo Dropout (Gal and Ghahramani, 2016, MC-DSim). We convert the scores into binary predictions by taking 10 equally-spaced cutoff values between the min and max scores of a method and compute precision, recall, and accuracy per cutoff.

Fig. 3 show that improving the precision of these heuristics requires sacrificing on recall. Accuracy scores in Tab. 1 highlight that while COMET often scores highest, it only improves marginally over



(a) Sum (b) DM

Figure 3: Selected examples of precision-recall curves

a majority baseline on Sum and DM. Focusing on fluent output has different effect across tasks: it benefits MT but lowers the scores for Sum and DM.[3] Also note the gap between source- and target-referential annotations, which leads to different optimal solutions for Sum (either Attn-ign-SRC or COMET). In all, detecting overgeneration across NLG tasks requires diverse methods and the best use of the model-aware setting is an open question.

5 Organizers

Elaine Zosa, Raúl Vázquez and Vincent Segonne have experience with seq2seq models and semantic annotation schemes (e.g. Martinc et al., 2022; Zosa et al., 2022; Barque et al., 2020; Raganato et al., 2021). Jörg Tiedemann has extensive experience with MT design and evaluation and organized WMT and VarDial shared tasks (Guillou et al., 2016; Zampieri et al., 2017). Alessandro Raganato and Timothee Mickus organized the SemEval 2023 Task 1 and 2022 Task 1 (Mickus et al., 2022; Raganato et al., 2023) and worked on seq2seq and annotation projects (e.g. Mickus et al., 2019; Raganato et al., 2020). Marianna Apidianaki was chair of SemEval from 2017 to 2019 and organized SemEval 2016 Task 5 (Pontiki et al., 2016).

[1] huggingface.co/Helsinki-NLP/opus-mt-mul-en

[2] huggingface.co/google/pegasus-xsum

[3] Remark it does not change which heuristic is most effective. None of the heuristics can detect fluency better than the majority baseline, justifying our treatment of disfluency as orthogonal to overgeneration.

6 Ethical Considerations

We strive to adhere to the ACL Code of Ethics in our work. [4]

Broader Impact Hallucinated outputs from large language models can be used to further spread disinformation and advance misleading narratives. Detecting hallucinated outputs is an important step in elucidating the factors of this phenomena and contribute to ongoing efforts to mitigate hallucination. This leads to the development of more trustworthy generative language models.

Data and Annotators Due to the nature of the proposed task, the data we release might contain false or misleading statements. In the case of annotated data, these statements will be labeled as such but not for the unannotated portions of the data.

Our annotators will be suitably compensated for their work and guaranteed a safe working environment. They will also be properly trained in the annotation task.

References

Lucie Barque, Pauline Haas, Richard Huyghe, Delphine Tribout, Marie Candito, Benoit Crabbé, and Vincent Segonne. 2020. FrSemCor: Annotating a French corpus with supersenses. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5912–5918, Marseille, France. European Language Resources Association.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. PROTAUGMENT: Unsupervised diverse short-texts paraphrasing for intent detection meta-learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2454–2466, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

[4] https://www.aclweb.org/portal/content/acl-code-ethics

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 525–542, Berlin, Germany. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Matej Martinc, Syrielle Montariol, Lidia Pivovarova, and Elaine Zosa. 2022. Effectiveness of data augmentation and pretraining for improving neural headline generation in low-resource settings. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3561–3570, Marseille, France. European Language Resources Association.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3259–3266. AAAI Press.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. Semeval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Alessandro Raganato, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. 2021. An empirical investigation of word alignment supervision for zero-shot multilingual neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8449–8456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabels participation in the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, 6(4):355–361.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Elaine Zosa, Lidia Pivovarova, Michele Boggia, and Sardana Ivanova. 2022. Multilingual topic labelling of news topics using ontological mapping. In *Advances in Information Retrieval*, pages 248–256, Cham. Springer International Publishing.