DeepPavlov at SemEval-2024 Task 6: Detection of Hallucinations and Overgeneration Mistakes with an Ensemble of Transformer-based Models

Ivan Maksimov and **Vasily Konovalov** and **Andrei Glinskii**
Moscow Institute of Physics and Technology
{maksimov.ivan.v, vasily.konovalov, glinsky}@phystech.edu

Abstract

The inclination of large language models (LLMs) to produce mistaken assertions, known as hallucinations, can be problematic. These hallucinations could potentially be harmful since sporadic factual inaccuracies within the generated text might be concealed by the overall coherence of the content, making it immensely challenging for users to identify them. The goal of the SHROOM shared-task is to detect grammatically sound outputs that contain incorrect or unsupported semantic information. Although there are a lot of existing hallucination detectors in generated AI content, we found out that pretrained Natural Language Inference (NLI) models yet exhibit success in detecting hallucinations. Moreover their ensemble outperforms more complicated models.

1 Introduction

Over the past few years, Natural Language Generation (NLG) models have experienced substantial advancements, particularly due to transformer-based architectures like a Generative Pretrained Transformer (GPT) (Radford et al., 2019). However, two interconnected issues challenge the field: firstly, the tendency of present neural systems to generate incorrect yet smooth outputs and, secondly, the inadequacy of existing metrics in evaluating accuracy over fluency. This causes NLG models to "hallucinate", i.e., produce fluent but incorrect outputs that we currently struggle to detect automatically (Ji et al., 2023).

The Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (SHROOM) has been suggested to address this challenge. In particular, the SHROOM task aims at addressing the existing gap in assessing the semantic correctness and meaningfulness of NLG models.[1] Within the Shared task (Mickus et al., 2024), one needs to detect grammatically sound English output that contains incorrect semantic information (i.e.,unsupported or inconsistent with the source input) in case there is no labeled training data available.

We propose to address the SHROOM task by leveraging an ensemble of pretrained transformer-based Natural Language Inference (NLI) models. The NLI models are used to derive features of hallucination probabilities, and then a tree-based gradient boosting model (Prokhorenkova et al., 2019) provides a final decision. Our results indicate that NLI-based models can be effectively used to detect hallucinations. Moreover, the ensemble model highly outperforms the base estimators in correlation with annotators' decisions.

To summarize, this work includes the following contributions:

- We conducted a systematic study, re-evaluating existing NLI models for hallucination detection tasks.
- We trained an ensemble of NLI models to detect hallucination that correlates with human judgment.

Additionally, we made the code publicly available.[2]

2 Background

Nowadays, it is well known that NLG models often generate coherent outputs that are not faithful to the given input, commonly referred as hallucinations (Maynez et al., 2020). Hallucination has been studied in a wide range of tasks, including but not limited to summarization (Huang et al., 2021), dialogue generation (Shuster et al., 2021) and a variety of other NLG tasks.

[1] https://helsinki-nlp.github.io/shroom/

[2] https://github.com/ivan-kud/semeval-2024-shroom

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 274–278

There are several benchmarks for hallucination detection. HaluEval includes 5,000 general user queries with ChatGPT responses and 30,000 task-specific examples from three tasks, i.e., question answering, knowledge-grounded dialogue, and text summarization (Li et al., 2023). FaithDial is a benchmark for hallucination-free dialogues by modifying hallucinated responses in the Wizard of Wikipedia (WoW) benchmarks (Dziri et al., 2022).

The SHROOM shared task organizers went one step further. The shared task was conducted with a newly constructed dataset of 4,000 model outputs labeled by five annotators each, including three NLP tasks: machine translation (MT), paraphrase generation (PG), and definition modeling (DM). Participants were asked to detect hallucinations in two different settings: a model-aware track where the organizers also provided a checkpoint to a model that generated the output and a model-agnostic track where they did not. The checkpoints are publicly available on HuggingFace.

All three NLG tasks are in English, with the exception of the input for the MT task, which is in Russian for the model-agnostic task and in many other languages for the model-aware task (Mickus et al., 2024).

3 Dataset

The dataset for the SHROOM challenge comprises a compilation of model-generated text entries with the aim to classify each output as either a hallucination of the generative model or not.

Information for the data sample includes the following fields: (i) *src* – the input text given to the generative language model; (ii) *hyp* – the generated textual output of the model; and (iii) *tgt* – the intended reference or the ground truth text that the model is supposed to generate; (iv) *task* – the task being solved; (v) *labels* – five labels, either *"Hallucination"* or *"Not Hallucination"* labeled by five annotators, and finally, (vi) *p(Hallucination)* indicates the proportion of annotators that labeled the data sample as a hallucination.

The dataset was split in the following way: training data of 30,000 samples without annotations with 10,000 samples for each task; validation data of 499 labeled samples with 187, 187, and 125 samples for DM, MT, and PG tasks, respectively; and test data of 1,500 examples without annotations to evaluate and rank the results of the competitors with 563, 562, and 375 examples for DM, MT, and PG tasks, respectively. Validation data sample is presented in Table 1.

All participants' submissions were evaluated using two criteria:

- Accuracy that the system reached on the binary classification.
- Spearman correlation of the system's output probabilities with the proportion of the annotators labeling the item as a hallucination.

4 Methodology

NLI task determines whether a hypothesis follows a premise and classifies it as either entailment, contradiction, or neutral. Previous research showed that NLI can be successfully used for hallucination detection in summarization (Gekhman et al., 2023).

Our system is an ensemble of several NLI pre-trained transformer-based models. Each model defines a new feature set, then a tree-based gradient boosting final estimator is trained on top of these features.

It's well known that in-domain training data are crucial for classifier performance (Konovalov et al., 2016). However, organizers do not provide us with the labeled training set, so we train final estimator on the labeled validation set.

The following is a list of NLI transformer-based models used in our pipeline:

- RoBERTa by Liu et al. (2019) is a robustly optimized BERT that outperformed BERT on the MNLI task.
- Sentence-RoBERTa by Reimers and Gurevych (2019) that returns sentence embeddings. Thus, we calculate the probability of hallucination as $1 - S$, where S is cosine similarity between the premise and hypothesis embeddings.
- DeBERTa by Manakul et al. (2023) uses an disentangled attention mechanism and an enhanced mask decoder being pre-trained on MNLI-m/mm outperformed BERT and RoBERTa of comparable sizes.
- T5 by Honovich et al. (2022) was trained on a binary problem (entailment or contradiction) on the bunch of NLU datasets: SNLI (Bowman et al., 2015), MNLI (Williams et al., 2017), Fever (Thorne et al., 2018), Scitail (Khot et al., 2018), PAWS (Zhang et al., 2019), VitaminC (Schuster et al., 2021).

Task	Input	Output	Reference	p(Halluc.)
DM	Because redpillers are usually normies or <define> Chadlites </define> while incels are ugly low - value males that women feel innately repulsed by .	(slang) An incel.	(incel, _, slang) A man of a slightly lower ranking on a scale of physical attractiveness than Chad.	0.2
MT	Ядрён батон!	Nuclear Baton!	Blimey!	1.0
PG	Are we resolved?	So, what, we're in the clear now?	Do we have a deal?	0.4

Table 1: Data examples from the validation sample of the model-agnostic track. Hallucination probability of 0.2 means that one of the annotators classified the example as a hallucination and the remaining four annotators classified it as not a hallucination. The resulting class label is determined by majority voting. For the DM task, the model had to define a word between two special tokens: <define> and </define>.

In addition to the output of the before-mentioned models, we add as features the lengths of input, output and reference texts. Then we train CatBoost (Prokhorenkova et al., 2019) models as metamodels on top of these features. Besides CatBoost model, we also train Random Forest (Breiman, 2001) implemented in scikit-learn library (Pedregosa et al., 2011) and LightGBM (Ke et al., 2017). CatBoost yields the best results among them.

5 Experimental setup

We do not use any preprocessing of input texts (premises and hypotheses). Neither do we use an unlabeled training set. So, the transformer-based models serve to obtain features from the validation and test sets, then the CatBoost metamodel is trained on the validation set and predicts the test set.

As for the CatBoost metamodel, we performed the following steps:

- We found the hyperparameters on the validation set by using Optuna (Akiba et al., 2019). Stratified k-fold cross-validation [3] with 10 splits was used for the classification model and k-fold cross-validation with 10 splits – for the regression model. The best parameters for the classification model for the model-agnostic task: `iterations = 216`, `learning_rate = 0.010`, `depth = 12`, and for the model-aware task: `iterations = 129`, `learning_rate = 0.005`, `depth = 9`. The best parameters for the regression model for the model-agnostic task: `iterations = 356`, `learning_rate = 0.029`, `depth = 5`, and for the model-aware task: `iterations = 317`, `learning_rate = 0.012`, `depth = 9`.
- We evaluated the metrics on the validation sample using repeated stratified k-fold cross-validation with 10 splits and 5 repeats.
- We trained it on the whole labeled validation set.
- We predicted test set labels.

6 Results

The results on the test set for both model-agnostic and model-aware tracks are presented in Table 2. There are scores for the baseline provided by organizers, best scores from the leader-board, individual transformer-based models and our system as a whole.

Among NLI pre-trained models, T5 model significantly outperformed other NLI models. However, our ensemble approach using features from all NLI pre-trained models significantly outperformed T5 in terms of correlation with annotators' decisions.

Our approach for model-agnostic case provided us with an accuracy of 82.1% and Spearman correlation of 0.752. With this approach, our team achieved the 6th place out of 41 in the competition for model-agnostic track. Only two teams achieved a higher Spearman correlation.

The same approach was applied for the model-aware track and provided us with an accuracy of 79.9%, which is the 8th place out of 38 in the com-

[3] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

Model	model-agnostic		model-aware	
	Accuracy	**Corr.**	**Accuracy**	**Corr.**
`nli-roberta-large`	62.8	0.608	66.2	0.566
`roberta-large-mnli`	73.7	0.611	73.0	0.549
`deberta-base-mnli`	72.8	0.617	73.1	0.597
`deberta-large-mnli`	75.7	0.701	75.5	0.688
`deberta-xlarge-mnli`	73.5	0.699	74.4	0.681
`deberta-v2-xlarge-mnli`	74.4	0.711	74.7	0.677
`deberta-v2-xxlarge-mnli`	76.1	0.729	75.9	0.691
`deberta-selfchecknli`	75.3	0.683	75.9	0.683
`t5_xxl_true_nli_mixture`	81.1	0.650	79.6	0.626
`baseline`	69.7	0.403	74.5	0.488
Our system$_{\text{submitted}}$	82.1	0.752	79.9	0.713
Our system$_{\text{best}}$	82.5	0.757	79.9	**0.722**
Best leaderboard	**84.7**	**0.770**	**81.3**	0.715

Table 2: The results of the accuracy and Spearman correlation metrics on the test sample for the model-agnostic and model-aware tracks.

petition. The value of Spearman correlation turned out to be 0.713.

More detailed results of the competition can be found in Mickus et al. (2024).

7 Conclusion

In this paper, we describe the ensemble system for hallucination detection by using transformer-based models. We present a simple, yet effective ensemble pipeline that provided us with results comparable with the best scores for the both tracks.

Future work might include thoughtful error analysis. Improved quality can be achieved by annotating unlabeled training set with LLMs (Ostyakova et al., 2023). In addition, a multilingual setup of NLI models can be used to develop multilingual hallucination detection system (Chizhikova et al., 2023; Konovalov et al., 2020). The proposed approach can be used standalone or can be integrated into the DeepPavlov framework (Burtsev et al., 2018).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

L. Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. Deeppavlov: An open source library for conversational ai. In *NIPS*.

Anastasia Chizhikova, Vasily Konovalov, and Mikhail Burtsev. 2023. Multilingual case-insensitive named entity recognition. In *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, pages 448–454, Cham. Springer International Publishing.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Vasily Konovalov, Pavel Gulyaev, Alexey Sorokin, Yury Kuratov, and Mikhail Burtsev. 2020. Exploring the bert cross-lingual transfer for reading comprehension. In *Computational Linguistics and Intellectual Technologies*, pages 445–453.

Vasily Konovalov, Oren Melamud, Ron Artstein, and Ido Dagan. 2016. Collecting Better Training Data using Biased Agent Policies in Negotiation Dialogues. In *Proceedings of WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies*, Los Angeles. Zerotype.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 Task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254, Prague, Czechia. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2019. Catboost: unbiased boosting with categorical features.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.