Team Bolaca at SemEval-2024 Task 6: Sentence-transformers are all you need

Béla Linus Rösener
Student/Uni-Tuebingen
bela.roesener@gmail.com

Ilinca Vandici
Student/Uni-Tuebingen
ilinca.vandici@uni-tuebingen.de

Hong-Bo Wei
Student/Uni-Tuebingen
hong-bo.wei@student.uni-tuebingen.de

Abstract

The prevalence of fluent over-generation hallucinations, grammatically correct but nonsensical text, poses a significant challenge to the reliability of Natural Language Processing (NLP) systems. These fabricated constructs, arising from factors like overfitting or data sparsity, can mislead users and undermine system efficacy. The SemEval-2024 Task 6, SHROOM, addresses this concern by offering a comprehensive evaluation platform. For our own contribution to the task we make use of a logistic regression classifier and a feed-forward ANN in order to provide a computationally economical, yet reliable solution to the the Task at hand.

1 Introduction

Fluent over-generation hallucinations, grammatically correct yet factually incorrect or contextually irrelevant text outputs, remain a persistent obstacle in NLP systems, particularly large language models (LLMs). Moreover, the coherent aspect of the output means that hallucinations are harder to detect than other types of erroneous generation,as discussed in (Guerreiro et al., 2022), particularly in tasks like machine translation, especially considering most metrics for measuring performance only account for fluency rather than correctness (Guerreiro et al., 2022) . In order to ensure that tools like LLMs, which are becoming increasingly popular among the general population, provide the user base with information that is faithful and coherent in the context of various language tasks, research in identifying instances of hallucinations has become necessary. In this paper, we present our contribution for the SemEval 2024 task [1], SHROOM, where we work on solutions for detecting and categorizing hallucinations, using the data made available for different types of language generation tasks, stemming from a model-aware and a model-agnostic track. Additionally, taking into account the fact that earlier studies have adopted a LLM-based few or zero shot learning approach to the problem, we opt for a computationally economical approach instead, using a two-pronged model making use of logistic regression and a simple feed-forward network.

2 Task Description

SHROOM challenges participants to develop a model-agnostic or model-aware binary classification system capable of identifying fluent overgeneration hallucinations in diverse NLP tasks like definition modeling, machine translation, and paraphrase generation.

The data consists of 61,080 text outputs, of which 1,080 are manually annotated instances (the rest being unlabeled).

3 Background

SHROOM represents a pivotal benchmark for advancing NLP systems' ability to discern and categorize fluent over-generation hallucinations. Its focus on real-world applicability through the model-agnostic track and its diverse dataset empower researchers to assess the limitations and strengths of current techniques. Ultimately, SHROOM contributes to the broader mission of enhancing the trustworthiness and resilience of NLP systems, a crucial aspect for applications like machine translation, text summarization, and chatbot interactions.

The task requires participants to develop a binary classification system which successfully identifies hallucinations for different types of language generation tasks: definition modeling, machine translation and paraphrase generation. The data was generated from two different tracks, model-aware, meaning knowlegde of the model which produced

[1]Our code is available for replication purposes at https://github.com/cicl-iscl/SemEval2024$_T$6$_S$$HROOMS$

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1663–1666

the output is accessible, and model-agnostic, where the model which generated the ouput is unknown. The generated outputs are provided in JSON format, containing source text, generated text and golden standard prompt, aswell the model name when applicable. 61080 datapoints are obtained in this way, from which 1080 are annotated (Mickus et al., 2024). A baseline was made available, using a zero-shot model with calls to LLAMA (Mickus et al., 2024).

4 Our System Strategy

Considering the likely instances where a system to detect hallucinations would find practical use and how current approaches to hallucination detection work (Friel and Sanyal, 2023), we decided that we wanted to make reduced inference time a goal of our system. Our primary strategy for detecting hallucinations involves a two-pronged approach utilizing either logistic regression or a small feed-forward neural networks trained on a labeled dataset as classification model. This leverages the strengths of each model for efficient and accurate hallucination detection. As input to our classification model we use sentence embeddings generated by SBERT (Reimers and Gurevych, 2019).

Using a logistic regression model as a baseline helps us to quantify the advantages of using a neural network as a classification model instead of simpler approaches.

4.0.1 SBERT

In order to enrich our understanding of the text and capture deeper semantic relationships beyond surface-level similarities, we incorporate Sentence-BERT (SBERT) embeddings into our system. SBERT generates high-dimensional vector representations of text, encoding semantic meaning and context.

We obtain reliable vector representations of our data utilizing a pre-trained SBERT model (e.g., all-mpnet-base-v2, all-MiniLM-L6-v2), we generate vector representations for each the source, target and hypothesis fields of our inputs. These vectors can be envisioned as high-dimensional fingerprints capturing the semantic essence of each sentence and its relationship to others. These SBERT-derived features are integrated with features such as task and model. This enriched feature set provides a comprehensive representation of the text, capturing deeper semantic information.

SBERT enables us to transcend basic word-level comparisons, allowing us to capture meaning and context within text more comprehensively. SBERT takes into account the context surrounding each sentence during analysis, which aids in identifying variations from the intended meaning and inconsistencies within the text. We anticipate that combining traditional features with those derived from SBERT will enhance the accuracy and generalizability of hallucination detection.

Additionally, using SBERT fits into our lightweight approach to the task, by offering a fast tool for inference, being more lightweight than newer state-of-the-art models.

4.0.2 Logistic Regression

For the initial layer of analysis, we employ logistic regression as a robust baseline model. Its interpretability allows us to gain insights into the key features distinguishing genuine and hallucinated text. We use SBERT to encode a prompt that incorporates Source, Target and Hypothesis. This provides the logistic regression model with single vector as input. The logistic regression model is then trained on these features to learn the underlying patterns that differentiate hallucinated and non-hallucinated text. This simple method provided us with a simple baseline.

4.0.3 Artificial Neural Network

The classification network is a simple multilayer feed-forward network. Its input are three sentence embeddings generated by SBERT from the Source, Target and Hypothesis fields of the input, as well as other features that the input provides. The usage of a neural network allows us to capture non-linear relationships and hidden patterns within the data that might be missed by the logistic regression model. The ANN architecture is designed with multiple hidden layers and non-linear activation functions, enabling it to learn intricate feature interactions and representations.

4.1 Key Discoveries and Challenges

The task presented us with two small datasets containing labeled instances and a bigger dataset without labels. Our main challenge therefore was to make the best use of a very limited dataset or find ways to leverage the not annotated data.

In regards to the following step in our approach, namely feature extraction, while many of the instances contained in the dataset seem to require

deep semantic analysis of the input data, much simpler features of the input can also be useful to identify hallucinations: word repetitions, n-gram counts, output length, unexpected characters, etc. (Huang et al., 2023)

Pertaining to supervised learning, while the approach proved itself to be useful, it also revealed its limitations. The reliance on pre-labeled data can restrict the generalizability of the model to new domains or tasks. Additionally, the quality of the pre-labeled data can impact the model's performance.

One of the main challenges we encountered was the lack of labeled data. Classifying hallucinations in unlabeled data remains a challenging task. The absence of explicit labels hinders the model's ability to definitively determine whether an instance is a hallucination. This problem highlights the need for more sophisticated methods for dealing with unlabeled data. However, the inclusion of the pre-labeled data which was made available for training our hallucination detection model proved to be an effective strategy. The model was able to generalize well to unseen data and achieve significant accuracy in identifying hallucinations, indicating that access to annotated datasets documenting cases of over generation will of course improve classification.

Our exploration of the dataset revealed that over-generation, the production of excessive or irrelevant text, is a significant aspect of hallucinations. The ability to distinguish between fluent over-generation and genuine hallucinations poses an additional challenge for our system, and is an aspect which would need to be further explored.

5 Key Algorithms and Modeling Decisions in Our SHROOM System

Our hallucination detection system employs a hybrid approach that combines both a pretrained LLM and a small classification model. The system's core components include:

1. Pre-trained Language Model (LM): We utilize a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to extract linguistic features from the text outputs ((Reimers and Gurevych, 2020)). BERT's ability to capture contextual information and semantic relationships is crucial for understanding the nuances of language and identifying deviations from the intended meaning. We use SBert for its ability to provide meaningful sentence embeddings and while being faster and less resource intensive than the newest LLMs openly available. This would make running our model in practical applications more realistic.

2. Classification Model: The sentence embeddings produced by the BERT-Model are given to 1) a logistic regression model or 2) a small feed-forward network producing a label probability.

3. Supervised Learning from Hallucination Annotations: The feed-forward network is trained using the labeled data provided in the SHROOM dataset. The model learns to classify text outputs as either containing hallucinations or being truthful to the Source.

6 Results

We achieved an accuracy of 0.57 on the model-aware track, and an accuracy of 0.63 on the model-agnostic track, placing us at respectively rank 32 and rank 27 on the competition leaderboard. Additionally, we scored 0.24 for accuracy for the model-agnostic track.

7 Experimental Setup

Because of our approach based on supervised learning we used the development dataset provided by the task organizers as training dataset, using cross validation to gain insights into our systems' performance before the actual test dataset was available.

We used PyTorch: 1.10.2 [2] in order to build our neural network. Transformers 4.12.2 [3], more specifically SBERT [4], was used in order to extract the sentence embeddings. Finally, in order to organize and process the data, we also made use of NumPy (1.22.3) [5] and Pandas (1.4.2)[6]. In order to run our code in an efficient manner, we used Colab[7].

8 Conclusion

By using a simple model exploiting feature extraction to aid in identifying hallucinations in a dataset containing data from different tasks, we

[2]https://pytorch.org/
[3]https://github.com/huggingface/transformers/
[4]https://www.sbert.net/
[5]https://numpy.org/
[6]https://pandas.pydata.org/
[7]https://colab.research.google.com/

have achieved an accuracy of 0.628. This could potentially indicate that for this type of task, it might be worthwhile to take into consideration approaches which do not exclusively rely on zero-shot classification, but instead make use of less computationally costly techniques. We have shown that such methods are not only efficient, but also present the advantage of being easily reproducible with fewer resources.

9 Going Forward

While our system is, at its current state, not usable in production systems, it shows that computationally less expensive methods can still lead to working systems in a task as complex as hallucination detection. Our implementation still leaves room for improvement and some unexplored possibilities: Our system does not leverage the unlabeled training data provided with the task. Using an encoder-decoder architecture to pretrain an encoder layer for the classification model might improve its training results on the small labeled data set. SBERT embeddings can be used to detect meaning similarities of texts, adding combinations of different embeddings (Source+Hypothesis, Target+Hypothesis, ...) may provide useful features to the classification layer. Our aim for a lightweight and fast system also makes manual approaches of hallucination detection as described in (Bruno et al., 2023) attractive.

References

Alessandro Bruno, Pier Luigi Mazzeo, Aladine Chetouani, Marouane Tliba, and Mohamed Amine Kerkouri. 2023. Insights into classifying and mitigating llms' hallucinations. *arXiv preprint arXiv:2311.08117*.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *arXiv preprint arXiv:2310.18344*.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *arXiv preprint arXiv:2208.05309*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.