Research Proposal

Michele Papucci

1 Research Context and Main Goal of the Project

Natural Language Processing (NLP) is a field of Computer Science and Linguistics that deals with the automatic analysis of human's natural language. While in the past decades, NLP problems were solved through the use of classifiers, today, the standard tools are Neural Language Models (NLMs), which are deep neural network architectures trained on massive amounts of data, that learn to extract dense representations of text, encoding syntactic and semantic information. These have reached new state-of-the-art performances even in hard text-generation tasks like summarization, text simplification, and translation.
However, these models still present some major problems, one being **hallucination**, which is defined as a generation that's *unfaithful* or *nonsensical* and presented as facts (Ji et al. 2023). It is especially dangerous since even when hallucinating, these models appear certain and fluent, giving no particular clues that what has been generated is wrong. This makes shipping these models to production, in moderate- to high-risk settings, dangerous or impossible (Lee et al. 2023).

The goal of the project is to use **Controlled Text Generation** (CTG) techniques (Zhang et al. 2023) to both guarantee that the model adheres to specified syntactic constraints (e.g. writing in a simpler form) and to identify and prevent hallucinations.

2 Detailed Description of the Project

The project aims to develop new techniques for CTG, controlling both the form and the content of the generation of a text-to-text NLM.
For syntax CTG, the idea is to focus on **text simplification tasks**, by implementing an approach inspired by Direct Preference Optimization (Rafailov et al. 2023), a Reinforcement Learning technique for CTG, that tunes the model generation towards desired outcomes using a preference dataset. This dataset can be built semi-automatically by scoring output generations, given a prompt, using well-established readability metrics such as READ-IT (Dell'Orletta et al. 2011), or CTAP (Chen et al. 2016). Then, the model's output can be automatically validated using linguistic metrics (Brunato, Cimino, et al. 2020), instead of relying on manual checks. This can be done thanks to the well-known correlation between some syntactic features and text complexity (Brunato, De Mattei,

et al. 2018), enabling us to guarantee the generations' simplicity. Next, the process will be extended to other syntax-dependent tasks, such as generating in a specific text genre or format (e.g. poetry) and generating text targeted for a certain level of education.

From the content point of view, the main challenge is to **automatically spot hallucinations**. Even building datasets to study them is hard and, indeed, we have few resources on them: Lin et al. 2022 built TruthfulQA, a dataset built with specially crafted questions that humans would answer falsely due to common misconceptions, but to use it as a benchmark, humans annotators are needed; Li et al. 2023 proposed HaluEval, a dataset of model generations annotated by humans for hallucinations. However, a classifier trained to spot hallucination using HaluEval performs decently on GPT-4 generations, and poorly on Alpaca's. This is due to the variation between hallucination type and scope between different models.

A promising model-free approach has been proposed in Sun et al. 2024, which presented a dataset containing unsolvable math problems. By parsing the outputs of the model on these problems, they were able to identify hallucinations. We'd like to extend the idea in a more textual scenario, creating a similar resource containing unsolvable textual questions, making it possible to parse the model outputs for identifying hallucinations. In certain text-to-text scenarios, such as text simplification and text summarization, a more algorithmic approach can be tested by generating from both the input and the output a Knowledge Graph (Hogan et al. 2021). By comparing these two graphs, we might find that certain differences are indicative of hallucination, creating a model-free approach for identifying them.

3 Impact

The proposed research aims to advance the understanding and capabilities of CTG techniques and hallucination identification, hopefully leading to the creation of better and more trustworthy tools. This could speed up the adoption of NLMs in high-stakes areas such as healthcare, finance, and education where accurate information is crucial. Also, the work on integrating linguistic measures for the validation of the NLMs outputs will enrich the scientific literature with new measures and algorithms for controlling these models.

References

Brunato, Dominique, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni (2020). "Profiling-ud: a tool for linguistic profiling of texts". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 7145–7151.

Brunato, Dominique, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, Giulia Venturi, et al. (2018). "Is this sentence difficult? do you agree?" In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. Association for Computational Linguistics, pp. 2690–2699.

Chen, Xiaobin and Detmar Meurers (Dec. 2016). "CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis". In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. Ed. by Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François, and Philippe Blache. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 113–119. URL: https://aclanthology.org/W16-4113.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi (July 2011). "READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification". In: *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*. Ed. by Norman Alm. Edinburgh, Scotland, UK: Association for Computational Linguistics, pp. 73–83. URL: https://aclanthology.org/W11-2308.

Hogan, Aidan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann (July 2021). "Knowledge Graphs". In: *ACM Computing Surveys* 54.4, pp. 1–37. ISSN: 1557-7341. DOI: 10.1145/3447772. URL: http://dx.doi.org/10.1145/3447772.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung (Mar. 2023). "Survey of Hallucination in Natural Language Generation". In: *ACM Comput. Surv.* 55.12. ISSN: 0360-0300. DOI: 10.1145/3571730. URL: https://doi.org/10.1145/3571730.

Lee, Peter, Sebastien Bubeck, and Joseph Petro (2023). "Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine". In: *New England Journal of Medicine* 388.13, pp. 1233–1239.

Li, Junyi, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen (Dec. 2023). "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 6449–6464. DOI: 10.18653/v1/2023.emnlp-main.397. URL: https://aclanthology.org/2023.emnlp-main.397.

Lin, Stephanie, Jacob Hilton, and Owain Evans (May 2022). “TruthfulQA: Measuring How Models Mimic Human Falsehoods”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229. URL: https://aclanthology.org/2022.acl-long.229.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn (2023). “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=HPuSIXJaa9.

Sun, Yuhong, Zhangyue Yin, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Hui Zhao (2024). *Benchmarking Hallucination in Large Language Models based on Unanswerable Math Word Problem*. arXiv: 2403.03558 [cs.CL].

Zhang, Hanqing, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song (Oct. 2023). “A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models”. In: *ACM Comput. Surv.* 56.3. ISSN: 0360-0300. DOI: 10.1145/3617680. URL: https://doi.org/10.1145/3617680.