Opportunity for a Ph.D. in Natural Language Processing:
Statistical Analyses of Lexical Distributions with an Application to Anomaly Detection in Natural Texts

July 6, 2023

Context and Motivation

Large Language Models (LLMs) a.k.a *foundation models* [2] have greatly improved the fluency and diversity of machine-generated text. Indeed, the release of ChatGPT and GPT-4 by OpenAI has sparked global discussions on the effective use of AI-based writing assistants. However, this progress has also introduced considerable threats such as fake news [1], and the potential for harmful outputs such as toxic or dishonest speech [4], among others. As it seems, the research on methods aimed at detecting the origin of a given text to mitigate the dissemination of forged contents and to prevent technology-aided plagiarism lags behind the rapid advancement of AI itself [10, 8]. For tasks like question-answering, it is essential to know when we can trust the natural language outputs of foundation models [7]. Likewise, for tasks like machine translation, it becomes important to detect hallucinations or omissions, i.e., translations that either contain information completely unrelated to the input or that do not include some of the input information [6, 5].

Recent works have indeed focused on tools that are able to spot such AI-generated outputs to identify and address these underlying risks. However, many of the existing approaches rely on pre-existing classifiers for specific undesired outputs, which restricts their applicability to situations where the harmful behavior is precisely known in advance.

Statistical analysis of lexical distributions is a valuable approach for anomaly detection in natural texts. By examining the frequency distributions of words and phrases in a given text or dataset, statistical methods can help identify unusual or anomalous patterns that deviate from the norm and these anomalies may indicate potentially harmful outputs, may reveal the origin of a given text, or detect hallucinations, stylistic inconsistencies, or even malicious intent in the text. By leveraging statistical methods to analyze lexical distributions, this thesis will focus on the automatic uncovering of deviations and anomalies that may indicate irregularities or unexpected patterns in natural language texts.

Our Exciting Opportunity

Forged texts and misinformation are ongoing issues and are in existence all around us in biased software that amplifies only our opinions for a “better”, more seamless user experience. On social media platforms, such software is used by rogue states, businesses, and individuals to create misinformation, amplify doubts about factual data or tarnish their competitors or adversaries, thereby enhancing their own strategic or economic positions. This spread may be the result of different factors and incentives; however, each poses the same fundamental issue to humanity: the misunderstanding of what is true and what is false.

Leveraging deep learning models for large-scale text generation such as GPT-3 and GPT-4 has seen widespread use in recent years due to superior performance over traditional generation methods, demonstrating an ability to produce texts of great quality, with a coherence and relevance that is sometimes hard to distinguish from human productions. These models generate text via an auto-regressive procedure that samples from a distribution learned to mimic the ”true” distribution of human written texts. Malicious uses of these technologies thus constitute a major threat to truthful information.

Artificial text detection can be viewed as a special case of anomaly detection, broadly defined as the task of identifying examples that deviate from regular ones to a degree that arouses suspicion. Current research in anomaly detection largely focuses either on deep classifiers (e.g., out-of-distribution detection [3], adversarial attack [9]) or relies on the output of large language models when labeled data is unavailable. Although these lines of research are appealing, they do not scale without requiring a large amount of computing. Additionally, these methods make the fundamental assumptions that (1) the statistical information needed to identify anomalies is available in the trained model, (2) the model uncertainty can be trusted, which is typically not the case as illustrated in the presence of a small shift in the input distribution. LLM-based approaches do not perform well when used on large text fragments, as may be needed in practical applications (e.g., novel, story, or news generation), because of the fixed length context used when training the language model.

Research Project

This Ph.D. thesis focuses on developing hybrid anomaly detection methods using deep neural network-based techniques and word frequency distributions that are linguistically inspired. Most of the research on language models to date focuses on sentence-level processing and fails to capture long-range dependencies at the discourse level. Instead, we will leverage word frequency distributions and information measures to characterize long documents, incorporating a very large number of rare words, which often leads to strange statistical phenomena such as mean frequencies that systematically keep changing as the number of observations is increased. Advanced concepts from statistics and information measures are necessary to understand the analysis of word frequency distributions and to capture document-level information. We are expected to design and develop novel statistical models and algorithms specifically tailored for analyzing lexical distributions in natural texts. Extensive experiments on real-world data sets will be executed to showcase the viability of our approach, benchmark its performance, and analyze its advantages, limitations, and areas for improvement.

Research questions. Some potential research questions for our consideration are:

- How can lexical distributions be effectively modeled and represented in natural language texts?
- What information (statistical) measures and techniques can be derived to identify anomalies in lexical distributions?
- How can contextual information and linguistic features be integrated into anomaly detection models based on lexical distributions?
- Can unsupervised learning techniques be leveraged to detect anomalies without the need for labeled anomaly data?
- How can domain-specific knowledge and expert (or mechanical) feedback be incorporated into the anomaly detection process to improve performance?

This research will provide a deeper understanding of statistical analysis techniques for anomaly detection in natural texts and contribute to the development of more accurate and reliable methods for identifying unusual patterns in language usage.

Team supervision. Institut des Systèmes Intelligents et de Robotique (ISIR) and the International Laboratory on Learning Systems (ILLS) are looking for a student with a background in AI and Data Science, who gets inspired by sciences and the opportunities of data and AI to solve complex NLP problems. You have strong programming skills and a very good understanding of data science, statistics, and Machine Learning.

An international and stimulating environment for research. ILLS will promote international mobility between France and Canada to facilitate collaborations with Ph.D students and professors in Canada. The university partners in Canada are: McGill University and École de Technologie Supérieure (ÉTS), and the Quebec Artificial Intelligence Institute (Mila), which are major players in AI at the international. They are involved in many research, industrial and academic projects. François Yvon, who will supervise this thesis at ISIR (Sorbonne Université), is a senior researcher at CNRS and a recognized expert in Automatic language processing, Machine translation, Speech recognition, Statistical language modeling, Document mining, Learning by analogy. Prof. Pablo Piantanida, who will supervise this thesis on the ILLS (McGill - ETS - Mila) side, is a recognized expert in information theory and Machine Learning. One of the strengths of the partners, is first the high level of the international within the recently created International Research Laboratory ILLS of the CNRS in Montreal, allowing a highly dynamic and rich research environment in AI at large.

Position Qualifications

- MSc program in Computer Science, Machine Learning, Computer Engineering, Mathematics, or related field (e.g. applied mathematics/statistics).
- Very good understanding of Machine Learning theory and techniques.
- Good programming skills in Python (PyTorch).
- Applications/ domain-knowledge in natural language processing is a plus.
- Good communication skills in written and spoken English.
- Creativity and ability to formulate problems and solve them independently.

How to apply. Applications should be sent by email to : francois.yvon@isir.upmc.fr; piantani@mila.quebec and also submitted via the link to the offer on the CNRS recruitment portal.

If you are interested, please send us the following elements as soon as possible and **not later than July 25th**:

- Detailed CV.
- Letter of motivation.
- Details of transcripts (especially M1 and M2).
- Elements of bibliography or personal achievements related to a research activity (e.g. master project, research internship subject, etc.).
- 2 letters of recommendation.

Applications with **missing elements** will not be considered.

References

[1] Rowan Zellers et al. "Defending Against Neural Fake News". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: `https://proceedings.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf`.

[2] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: `2108.07258 [cs.LG]`.

[3] Eduardo Dadalto Camara Gomes et al. "Igeood: An Information Geometry Approach to Out-of-Distribution Detection". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=mfwdY3U_9ea.

[4] Stephen Casper et al. *Explore, Establish, Exploit: Red Teaming Language Models from Scratch*. 2023. arXiv: 2306.09442 [cs.CL].

[5] David Dale et al. *HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation*. 2023. arXiv: 2305.11746 [cs.CL].

[6] Nuno M. Guerreiro et al. *Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation*. 2023. arXiv: 2212.09631 [cs.CL].

[7] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation*. 2023. arXiv: 2302.09664 [cs.CL].

[8] Eric Mitchell et al. *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*. 2023. arXiv: 2301.11305 [cs.CL].

[9] Marine Picot et al. "Adversarial Robustness Via Fisher-Rao Regularization". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (Mar. 2023), pp. 2698–2710. DOI: 10.1109/tpami.2022.3174724. URL: https://doi.org/10.1109%2Ftpami.2022.3174724.

[10] Xianjun Yang et al. *DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text*. 2023. arXiv: 2305.17359 [cs.CL].