ELOQUENT Lab @ CLEF 2024

Voight-Kampff task

This task is intended to address the general question

"Can text authored by generative language models be distinguished from text written by human authors?"

and the companion question

"Can text authored by some specific model be identified as such?"

Extended deadline: Submit your experiments by May 9 AoE!

Motivation

Recent advances in generative language models have made it possible to automatically generate content for websites, news articles, social media, etc. The EU has recently suggested to technology companies that labelling such AI-generated content as such might be a useful tool to combat misinformation and to protect consumer rights.

Goals of the task

This task will explore whether automatically-generated text can be distinguished from human-authored text. Detecting automatically generated text, with the increased quality of generative AI, is becoming a task quite similar to human authorship verification and this task will be organised in collaboration with the PAN lab with years of experience from previous shared tasks on authorship verification and closely related tasks.

This task will also investigate if models can be self-assessed in a reliable way with minimal human effort.

Procedure

1. Organisers pick a number of human-authored texts of about 500 words in genres such as:
    - Newswire
    - Wikipedia intro texts
    - Fan fiction
    - Biographies
    - Weather and stock market reports; sports results
    - Podcast transcripts

2. Descriptions for those texts are generated automatically
    - Bullet points, e.g.
    - Should capture genre and some of stylistic characteristics of the original
    - Maybe use datasets for summarisation evaluation

3. Ask participants to use their systems generate a text from those descriptions

4. Pass the resulting sets to PAN builders to see if
    - human texts can be distinguished
    - system characteristics can be tracked across texts, i.e. if the output of a system is similar enough across texts and genres to be identifiable

Data

The data are distributed in a format which resembles standard benchmark tests. The test collection has a suggested **prompt** string, and a list of items with a **Content** and an optional **Genre and Style** field. These can be used together with the suggested prompt string to generate a text. If participants wish to change the suggested prompt string to something more suitable, this is allowed, but this must be reported upon submission and in the written report of the experiment.

Task 4 sample and test topics on Huggingface

Submission format

The submission should be in a plain zip file of a directory named after the team with the generated texts in plain text form:

OurTeamName/006.txt

…

OurTeamName/029.txt

Submit your generated texts through this form

Extended deadline: Submit your experiments by May 9 AoE!

Result scoring

System outputs are scored by how often they fool a classifier into believing the output was human-authored.

Example

Below an original and a GPT-3.5-generated version for the one of the sample summaries. Can you tell which one is the original?

Text sample 1	Text sample 2
The Uralic languages, spoken by approximately 25 million people, have a rich linguistic history that dates back between 7,000 to 10,000 years ago. These languages are predominantly found in northeastern Europe, northern Asia, and North America. Hungarian, Estonian, and Finnish stand out as the most significant among the Uralic languages. However, attempts to trace their genealogy to earlier periods have been difficult due to the lack of concrete evidence. Nonetheless, there exists speculation regarding the relationship between Uralic and Indo-European languages, although they are generally not thought to be related. The Uralic languages can be divided into two main groups: Finno-Ugric and Samoyedic.	Uralic languages, family of more than 20 related languages, all descended from a Proto-Uralic language that existed 7,000 to 10,000 years ago. At its earliest stages, Uralic most probably included the ancestors of the Yukaghir language. The Uralic languages are spoken by more than 25 million people scattered throughout northeastern Europe, northern Asia, and (through immigration) North America. The most demographically important Uralic language is Hungarian, the official language of Hungary. Two other Uralic languages, Estonian (the official language of Estonia) and Finnish (one of two national languages of Finland—the other is Swedish, a Germanic language), are also spoken by millions. Attempts to trace the genealogy of the Uralic languages to periods earlier than

Both of these groups have given rise to various subgroups of languages, displaying their own unique characteristics and dialects.

In terms of similarity within the Finno-Ugric languages, it is comparable to the similarity between English and Russian. This highlights the significant diversity found within this group. For instance, Finnish and Estonian, which are closely related members of the Finno-Ugric subgroup, differ similarly to diverse dialects of the same language.

Overall, the Uralic languages have a fascinating history that can be traced back thousands of years. Although efforts to trace their genealogy to earlier periods have been hindered by limited evidence, it is evident that these languages have flourished within northeastern Europe, northern Asia, and North America. The prominence of Hungarian, Estonian, and Finnish showcases the significance of the Uralic languages within these regions.

While Uralic and Indo-European languages are generally not believed to be related, there is still room for speculation. The existence of similarities and shared features among these language families raises interesting questions about their potential connections.

The division of Uralic languages into Finno-Ugric and Samoyedic highlights the distinct linguistic developments within each group. These groups have given rise to numerous subgroups, each with its own set of languages and dialects. This diversity within the Uralic language family is comparable to the varied dialects within a single language, such as English or Russian.

In conclusion, the Uralic languages have a complex and ancient history, dating back thousands of years. Despite the challenges in tracing their genealogy, these languages have thrived in northeastern Europe, northern Asia, and North America. The significant Uralic languages, namely Hungarian, Estonian, and Finnish, demonstrate the importance and diversity within this language family. The existence of Finno-Ugric and Samoyedic groups further highlights the rich

Proto-Uralic have been hampered by the great changes in the attested languages, which preserve relatively few features and therefore provide little evidence upon which scholars may base meaningful claims for a more distant relationship. Most commonly mentioned in this respect is a putative connection with the Altaic language family (including Turkic and Mongolian). This hypothetical language group, called Ural-Altaic, is not considered by most scholars to be soundly based.

Although the Uralic and Indo-European languages are not generally thought to be related, more speculative studies have suggested a connection between them. Relationship with the Eskimo languages, Dravidian (e.g., Telugu), Japanese, Korean, and various American Indian groups has also been proposed. The most radical of these claims is the massive Dené-Finnish grouping of Morris Swadesh, which encompasses, among others, Sino-Tibetan (e.g., Chinese) and Athabaskan (e.g., Navajo).

The Uralic language family in its current status consists of two related groups of languages, the Finno-Ugric and the Samoyedic, both of which developed from a common ancestor, called Proto-Uralic, that was spoken 7,000 to 10,000 years ago in the general area of the north-central Ural Mountains. At its very earliest stages Uralic most probably included the ancestors of the Yukaghir languages (formerly listed as a Paleo-Siberian stock with no known relatives).

Over the millennia, both Finno-Ugric and Samoyedic branches of Uralic have given rise to more or less divergent subgroups of languages, which nonetheless have retained certain traits from their common source. For example, the degree of similarity between two of the least closely related members of the Finno-Ugric group, Hungarian and Finnish, is comparable to that between English and Russian (which belong to the Indo-European family of languages). The difference between any Finno-Ugric language and any Samoyedic tongue would be even greater. On the other hand, more closely related members of

linguistic developments that have taken place within the Uralic languages.	Finno-Ugric, such as Finnish and Estonian, differ in much the same manner as greatly diverse dialects of the same language.

Bibliography

- Yu, Peipeng, Jiahan Chen, Xuan Feng, and Zhihua Xia. "CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts." https://arxiv.org/pdf/2304.12008.pdf
- Hans, Abhimanyu, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. "Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text." (2024) https://arxiv.org/pdf/2401.12070.pdf
- Holly Else. 2023. Abstracts written by chatgpt fool scientists. Nature, 613(7944):423–423.
- Leon Fröhling and Arkaitz Zubiaga. 2021. Featurebased detection of automated language models: tackling gpt-2, gpt-3 and grover. PeerJ Computer Science, 7:e443.
- Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2022. Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv, pages 2022–12. Sebastian Gehrmann, SEAS Harvard, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. ACL 2019, page 111.
- Daphne Ippolito, Daniel Duckworth, Chris CallisonBurch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1808–1822.
- Mohammad Khalil and Erkan Er. 2023. Will chatgpt get you caught? rethinking of plagiarism detection. arXiv preprint arXiv:2302.04335.
- Gabriel Levin, Raanan Meyer, Eva Kadoch, and Yoav Brezinov. 2023. Identifying chatgpt-written obgyn abstracts using a simple tool. American Journal of Obstetrics & Gynecology MFM, 5(6).
- Sandra Mitrovic, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. arXiv preprint arXiv:2301.13852.
- Kimberley Mok. 2023. Chatgpt writes scientific abstracts that can fool experts. https://thenewstack.io/chatgpt-writes-scientific-abstracts-well-enough-to-fool-experts/.
- H. Holden Thorp. 2023. Chatgpt is fun, but not an author. Science, 379(6630):313–313.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, pages 9054–9065.
- ZeroGPT. 2023. Ai text detector. https://www.zero gpt.com.
- Mitrović S, Andreoletti D, Ayoub O. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text[J]. arXiv preprint arXiv:2301.13852, 2023.
- Guo B, Zhang X, Wang Z, et al. How close is chatgpt to human experts? comparison corpus, evaluation, and detection[J]. arXiv preprint arXiv:2301.07597, 2023.
- Areg Mikael Sarvazyan, José Ángel González, Paolo Rosso, and Marc Franco-Salvador. Supervised machine-generated text detectors: Family and scale matters. In Experimental IR Meets Multilinguality, Multimodality, and Interaction — 14th International Conference of the CLEF Association. 2023.