Limitations of Human Identification of Automatically Generated Text

Nadége Alavoine, Maximin Coavoux, Emmanuelle Esperança-Rodier, Romane Gallienne, Carlos-Emiliano González-Gallardo, Jérôme Goulian, Jose G Moreno, Aurélie Névéol, Didier Schwab, Vincent Segonne, et al.

Limitations of Human Identification of Automatically Generated Text

Nadége Alavoine[1], Maximin Coavoux[2], Emmanuelle Esperança-Rodier[2], Romane Gallienne[3], Carlos-Emiliano González-Gallardo[4], Jérôme Goulian[2], Jose G. Moreno[5], Aurélie Névéol[6], Didier Schwab[2], Vincent Segonne[7] and Johanna Simoens[8]

[1]Université Paris-Saclay, LISN, Campus Universitaire bâtiment 507, Rue du Belvédère, 91400 Orsay, France
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[3]Université Sorbonne Nouvelle, Lattice, CNRS, ENS-PSL, 1 rue Maurice Arnoux, 92120 Montrouge, France
[4]La Rochelle Université, L3i, 17000 La Rochelle, France
[5]University of Toulouse, IRIT, 31000 Toulouse, France
[6]LISN, Université Paris-Saclay, CNRS, 91403 Orsay, France
[7]Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
[8]Everteam, Bagneux, France
nadege.alavoine@universite-paris-saclay.fr, {first.last}@univ-grenoble-alpes.fr
romane.gallienne@cnrs.fr, carlos.gonzalez_gallardo@univ-lr.fr,
jose.moreno@irit.fr, aurelie.neveol@lisn.upsaclay.fr
vincent.segonne@univ-ubs.fr, johanna.simoens@gmail.com

Abstract

Neural text generation is receiving broad attention with the publication of new tools such as ChatGPT. The main reason for that is that the achieved quality of the generated text may be attributed to a human writer by the naked eye of a human evaluator. In this paper, we propose a new corpus addressing computer science topics in French and English for the task of recognising automatically generated texts and we conduct a study of how humans perceive the text. Our results show, as previous work before the ChatGPT era, that the generated texts share some common characteristics with human texts but they are not clearly identifiable which impacts the perception of synthetic texts.

Keywords: human identification, neural text generation, ChatGPT

1. Introduction

Human annotations are considered gold standard labels for multiple natural language processing (NLP) tasks such as morphological analysis, syntactic parsing, and lexical and relational semantics, to mention a few. Recent efforts in NLP focus on the use of neural models to fit the labels assigned by a human to a given text. However, as the quality of the system response becomes stronger, it becomes harder to know if a given prediction is the result of a human annotation or just the output of a system. For example, in machine translation, Wintner (2016) showed that humans struggle to identify whether a text is native or translated, while classification models achieve high performance. Similarly, Ippolito et al. (2020) showed that, when Large Language Models (LLM) generate the text, humans can strive to detect if the text was automatically generated.

As many setups may involve text generation, in this paper, we identified three main setups when automatically generating text and depicted them in Figure 1. The "Isolate" setup refers to the use of prompts that are included in the system answer while "InContext" makes explicit the interaction between two parties. Finally, the "InConversation" setup is similar to "InContext" but with multiple turns. Although it is required to deal with all of them nowadays, in this work we focused on the "InContext" setup. In this setup, for texts generated by low-quality tools, a superficial reading may be sufficient for a human to perform the authorship attribution task. However, as shown in this work, the recent ChatGPT model generates fluent text, which requires much closer inspection.

The contributions of this paper are threefold. First, we construct a dataset in French and English for the authorship attribution task. Second, we conduct an annotation campaign in which human participants undertook the text authorship attribution task. Third, we analyse the annotations from the campaign in order to characterise the obtained predictions. Our full corpus is publicly accessible at https://zenodo.org/records/10853531.

2. Background and Related Work

Although the evaluation of automatically generated text has been a large source of research interest, the recent public access to large conversational agents like ChatGPT has increased and enlarged the research community. Within this context, we intend to focus on automatically generated text de-

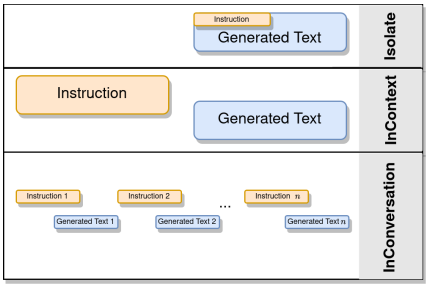

Figure 1: Setups of automatically generated texts

tection of works on the ChatGPT level. Thus, we reviewed the existing datasets for this task.

Related studies Evaluating if a text is automatically generated or not is presented by Uchendu et al. (2020) as a kind of Turing test. In the context of English news content, they evaluated whether a model was able to detect generated news in balanced or imbalanced setups. As a main conclusion, the problem is considered unsolved for texts generated with models such as GPT-2 (Solaiman et al., 2019). However, the best-performing model – a finetuned RoBERTa (Liu et al., 2019) – was able to achieve 0.98 accuracy on the task. Ippolito et al. (2020) evaluated the capacity of students to discriminate between human-written and automatically generated text. Although they used state-of-the-art systems at that time (GPT-2), the students were able to achieve between 64% and 77% accuracy over 475 responses.

Finally, the most similar task to our work is Autextification[1] subtask 1 - Human or Generated. This is a study organised as an evaluation task when participants are requested to submit system outputs to detect whether a text is automatically generated or not. However, the focus is on the system's capabilities rather than on the human's capabilities to address the problem (Sarvazyan et al., 2023).

Related datasets Current available datasets are mainly based on GPT and GPT-2 generated content as collecting outputs from APIs and humans may be complex and expensive. The *Human ChatGPT Comparison Corpus (HC3)* (Guo et al., 2023) is a recent dataset based on ChatGPT composed of 12,853 and 24,332 questions in English and Chinese respectively. This dataset is based on multiple sources including WikiQA, Wikipedia, and LegalQA, to mention a few. Despite the new production of the ChatGPT-based corpus, to the best of our knowledge, there is no corpus with a focus on the produced labels by humans after the collection from ChatGPT.

[1] sites.google.com/view/autextification

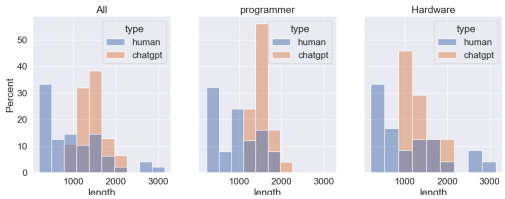

Figure 2: Length distribution between human and machine-generated answers

3. Our corpus

In this section, we describe how our corpus has been collected and compiled. As ChatGPT is a chatbot that can correct itself, we decided to see if its answers to a specific question could be distinguished from the answers given by a human being. All this is in the "InContext" setup as presented in Figure 1. To be able to distinguish between human and non-human answers to questions, we needed to find questions whose related answers were attested to be given by a human being. We thus decided to use the most common Q&A platforms, which are Quora in French[2] and Stackoverflow in English [3] to create the two corpora, one in French and the other one in English.

3.1. French corpus

We used the French version of Quora to collect questions and human answers. We selected the "*materiel informatique*' '(*hardware*) topic as well as the "*programmeur*" (*programmer*) topic. We set up a list of five criteria to select questions:

1. Recent questions: we filter out old questions that may be included in the ChatGPT training corpus, based on their publication date.
2. Questions with at least one answer: questions without any answer were ignored.
3. Not a translation: Quora allows users to visualise questions in other languages by automatically translating them. We only keep questions originally written in French.
4. Short length: for the sake of simplicity, we kept answers whose size was relatively short to avoid any length limitation when collecting the ChatGPT answers.
5. Only text, no images or other content: we ignored any questions when visual content or links were needed to provide an answer.

We decided to look for questions posted during the last quarter of 2022 and the first quarter of 2023, to make sure that the developers of ChatGPT did not include those questions in the training data.

[2] fr.quora.com/
[3] stackoverflow.com/

We ensured that there was an answer that would be considered as our human answer. We also checked that the answer did not correspond to a translation, as we noticed that in several cases the French answers were translations of answers to the same questions asked in English. It was indicated in the Quora interface as "This response may not be a faithful translation of the response from *User* in Quora in English: *Original Question*".[4] Eventually, we selected questions for which no additional content was needed to understand what was asked. For example, when a screenshot was added to the query showing what the problem was related to, we excluded this query from our corpus as it might interfere with the production of an answer. Once we had selected the questions as well as the answers from Quora, we copy-pasted the questions as they stood in a ChatGPT prompt. Then, fully generated answers were added to our corpus. We did this manually as Quora policy strictly restrict automated use.[5] ChatGPT responses have thus been collected for the entire dataset of 49 questions. Figure 2 shows the length distribution of our French dataset.

3.2. English corpus

Turning to the English corpus, we decided to collect questions and answers from Stackoverflow. We used the same list of criteria as the one listed in the previous subsection but item 3, as the issue with translated answers did not occur. Additionally, item 2 was extended to ensure that the answer used for the human is the one selected by the user who posted the question. Nevertheless, we decided to collect questions that were issued during the last weeks of December 2022. Furthermore, at least one answer selected by the user who asked the question was considered as the human answer. Since Stackoverflow does not have any policy on the use of API to collect the queries and their answers we did this automatically. We followed the same process as the one described in the above subsection to get ChatGPT responses for the entire dataset to end up with a total of 145 questions.

3.3. Annotation

Annotations were collected during the hOUPSh 2023 workshop,[6] that was co-located with TALN 2023, a NLP conference held in Paris in June 2023. We gathered all annotators (participants of the workshop, some of whom also coauthored this paper) into a single room to present the goal of the annotation task. As most of the participants were French

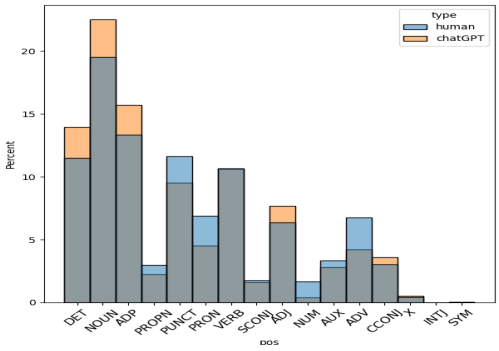

Figure 3: Distribution of POS categories in Human vs. ChatGPT answers

speakers, we mainly focused on the French corpora to be annotated during the event and made the English corpus available for interested participants to continue annotations after the event. Annotations were collected using DOCCANO[7] which is an open-source text annotation tool for humans. We divided the corpus into two test sets. For the first set, we provided questions and their corresponding answers from ChatGPT or the human answer from Quora. That is to say that the questions were only seen once by the participants. For the second test set, we provided the questions with the related human answers from Quora and the same questions with ChatGPT answers. In this case, the participants were presented twice with the same questions but each time randomly with a different answer from a human being or ChatGPT. We had 17 participants, all working in the field of NLP, most of them in academia (professors or PhD students). For both sets, they had to decide if the answer was created by a human being (*humain*) or by ChatGPT (*non humain*).

4. Results and Analysis

4.1. Human vs. ChatGPT answers

Since our corpus gathers answers both from humans and ChatGPT, we were able to perform some descriptive analysis and investigate some linguistic features to see how different the answers were.

Answers' length distribution We first observed the distribution of lengths of the answers and noticed a clear distinct pattern between human and ChatGPT answers. Indeed, as shown in Figure 2, humans tend to answer more concisely with 30% of the answers' length shorter than 500 characters while the distribution of ChatGPT's answer lengths are much more uniform and range essentially between 1,000 and 1,500 characters. This uniformity is recurrent in ChatGPT answers.

[4] Translated to English for the paper.

[5] In §4.4 www.quora.com/about/tos

[6] https://houpsh2023.sciencesconf.org/

[7] github.com/doccano/doccano

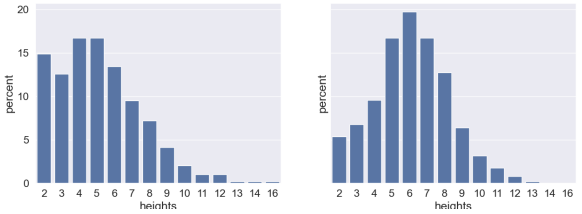

Figure 4: Distribution of tree heights for answers

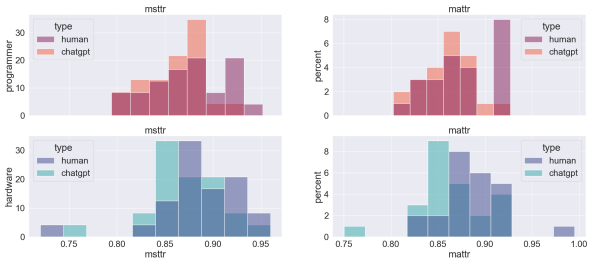

Figure 5: MSTTR and MATTR scores on human and ChatGPT answers

Syntactic features We analyse distributions regarding two different features which obtained using SpaCy's *fr_core_news_lg* model[8]. First, we considered the frequency of parts of speech in the answer tokens. Figure 3 shows the distributions of human and ChatGPT answers. From this figure, we can see that no strong differences can be drawn in the class distribution. However, when seen at tree height, the ChatGPT answers have deeper syntactic trees as shown in Figure 4.

Lexical richness As in Machine Translation (MT), lexical richness (Vanmassenhove et al., 2019) has shown that MT outputs are lexically less diverse than human translation. We aimed to see if we could notice such a phenomenon with automatically generated texts. Thus, we used the LexicalRichness[9] Python module to obtain such metrics. As the Type-Token Ratio (TTR) is influenced by sentence length, we computed the Mean Segmental TTR (MSTTR, Johnson, 1944) and the Moving Average TTR (MATTR, Covington and McFall, 2010), as they seem to be better indicators than TTR (Tezcan et al., 2019). MATTR is also more informative than MSTTR which indicates trends. Figure 5 shows the MSTTR and MATTR for human and ChatGPT answers according to the topics, i.e. *hardware* and *programmer*. We see that the distributions of MSTTR for human and ChatGPT answers overlap. Even if ChatGPT has a high lexical richness, the MSTTR distribution shows that human answers are more lexically diverse than ChatGPT ones, for both topics. This is confirmed by the MATTR distribution.

[8] spacy.io/models/fr
[9] github.com/LSYS/LexicalRichness

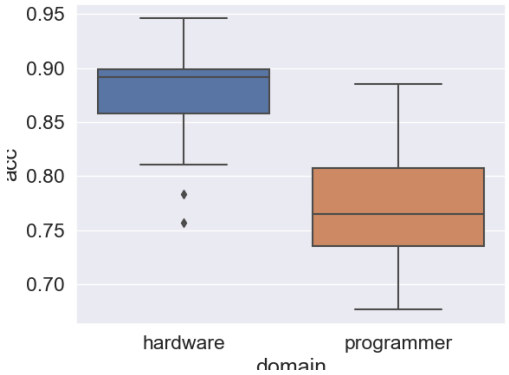

Figure 6: Distribution of the human annotation accuracy.

4.2. Annotation results

Inter-annotator agreement We measured the inter-annotator agreement using the Fleiss-kappa score (Fleiss, 1971, hereafter denoted FKS) which is an adapted version of the original Cohen kappa score (Cohen, 1968) for many annotators. As shown in Table 1, the overall inter-annotator agreement is 0.57 which is considered a medium agreement. In most cases, the annotators would predict the same label (human or ChatGPT) but we observe some disparities in specific cases. For example, it appears that annotators had higher agreement on answers drawn from the *hardware* topic (FKS=0.69) than from the *programmer* topic (FKS=0.47).

Human annotations accuracy We now turn to the evaluation of the annotators' performance regarding their ability to detect artificially generated answers. To do so, we measure the accuracy of the annotations and report the results in Table 2. On average, the annotators were able to distinguish human answers from ChatGPT answers 81% of the time overall. Having a closer look at the accuracy per topic of the answers, we notice that the gap previously observed in the inter-annotator agreement scores is also found in the accuracies as shown in Figure 6. This highlights even more the fact that ChatGPT answers are very close in style to human ones in the *programmer* topic.

Group	Hardware	Programmer	$All_{hardware+programmer}$		
			1 answer	2 answers	$All_{1+2answers}$
GR1	0.78	0.49	0.62	0.63	0.62
GR2	0.61	0.45	0.46	0.56	0.52
$All_{gr1+gr2}$	0.69	0.47	0.53	0.57	0.57

Table 1: Fleiss-kappa scores per group

Annotator	Overall	Programmer	Hardware
Human	0.81	0.77	0.87
ChatGPTDetector	0.86	0.76	0.95

Table 2: Accuracy of human and ChatGPTDetector annotations

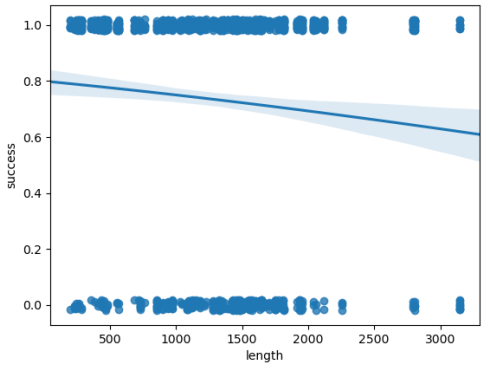

Figure 7: Average of the human annotation accuracy.

Automatically detect ChatGPT We also evaluated an automatic system to detect ChatGPT (Antoun et al., 2023), referred to as *ChatGPT Detector*, and compared its performance to the human annotations. The results are presented in Table 2 and show that the automatic system outperforms humans overall in discriminating humans' answers from ChatGPT ones. Interestingly enough, it follows the same behaviour as humans regarding the performance per topic. Indeed, while ChatGPT Detector outperforms humans by a large margin on the *hardware* topic, it encounters the same difficulty on the *programmer* topic and loses almost 20 points absolute in accuracy score.

To fairly compare the preferences of the participants we split them into two groups. Each group covered 75% of the data to annotated with only 50% of the data being annotated by both of them. This allows us to compare preferences in equal data and different data. Figure 6 shows the accuracy distribution between the two distributions. From this figure, we can see that: (1) both groups behave similarly in common questions but (2) behave strongly differently when different questions are annotated.

4.3. Length impact

We analyse the impact of the size on the annotators' performance. Figure 7 shows the average performance (curve) as well as individual performances (dots) concerning the answer size. From the figure, we conclude that, on average, shorter answers are easier to classify than larger ones.

5. Conclusions

This paper presents a new dataset for the task of recognising if a text was automatically generated or not. Additionally, we annotated it with quality annotations collected from humans exclusively and analysed them. Results indicate that the task is easier for humans when the texts are short, but the complexity of the task increases in larger texts. We explored multiple aspects to characterise this difference in the perception of the texts. However, further experiments and analyses are needed to clearly identify the reason for this observation.

6. Limitations

Despite our efforts to have high-quality annotated data, external factors may influence the indicated preferences by the users. A larger annotated dataset may reduce the impact of these factors.

- We compare a single answer provider (ChatGPT) to multiple ones (humans).
- Although we made an effort regarding the date of the posted content, we can not assert that a human answer has actually been provided by a human and not by an automatic system as humans could easily have copied/pasted answers from an automatic text generator.

Acknowledgements

This work has been supported by the *collège Technologies du Langage Humain* (TLH) of the *Association française pour l'Intelligence Artificielle* (AFIA). We also acknowledge the ANNA (2019-1R40226), TERMITRAD (2020-2019-8510010), Pypa (AAPR2021-2021-12263410), Actuadata (AAPR2022-2021-17014610) projects funded by the Nouvelle-Aquitaine Region (France), as well as the *Association pour le Traitement Automatique des Langues* (ATALA) and the *Association Francophone de Recherche d'Information* (ARIA) for hosting the conference (CORIA-TALN-2023) where the annotations were collected.

7. Bibliographical References

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model-generated text: Is ChatGPT that easy to detect? In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 14–27, Paris, France. ATALA.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Michael Covington and Joe McFall. 2010. Cutting the gordian knot: The moving-average type-token ratio (mattr). *Journal of Quantitative Linguistics*, 17:94–100.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, 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, Online. Association for Computational Linguistics.

Wendell Johnson. 1944. Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.

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*.

Areg Mikael Sarvazyan, José Ángel González, Marc Franco-Salvador, Francisco Rangel, Berta Chulvi, and Paolo Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Arda Tezcan, Joke Daems, and Lieve Macken. 2019. When a 'sport' is a person and other issues for NMT of novels. In *Proceedings of the Qualities of Literary Machine Translation*, pages 40–49, Dublin, Ireland. European Association for Machine Translation.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8384–8395.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Shuly Wintner. 2016. Translationese: Between human and machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 18–19, Osaka, Japan. The COLING 2016 Organizing Committee.