Natural Language Processing Subfield Survey: Detection of Computer-Generated Text

Shreya Nagpal

Abstract

In a world where generative language models like ChatGPT are becoming increasingly common and integrating into people's daily lives, misinformation and plagiarism are becoming more and more widespread. The study of methods to detect computer-generated text has been ongoing for decades, with early works focusing on detection of machine-translated text and more recent works focusing on identifying text generated by advanced models like Generative Pre-trained Transformers (GPT) and neural networks.

1 Introduction [1]

Computer generated text refers to any text created by a computer, generally after being prompted by a human being. Among other sources, this includes text generated by newly mainstream natural language models such as GPT and text generated by machine translation tools (Gatt and Krahmer, 2018). The recent boom in GPT-3 and chatbots that use predictive language to respond to human queries is an extremely useful tool, but it can also be used to spread misinformation. Tools such as ChatGPT are very good at generating natural-sounding text, but accuracy of information generated varies greatly – combined, these factors result in unreliable text that human beings believe easily (Bin Guo, September 2019). Additionally, such tools can be used for plagiarism and academic dishonesty with low likelihood of detection with similarity checkers (Michael Jones, January 2014). With artificial intelligence making misinformation straightforward to spread and dishonesty an unbelievably easy option, computer-generated text detection becomes the obvious field to focus on. How can one identify if a given piece of text is computer generated? How confident can one be in this prediction?

[1] Word Count: 2196

As machine-generated text becomes increasingly difficult to tell apart from human-generated text, the authorship attribution problem becomes a centerpiece in deciding whether or not a certain text is machine-generated. This problem has three forms:

- Determining if two texts are both human-generated or both machine-generated.
- Determining if any single given text is human-generated or machine-generated.
- Determining which neural method generated a text given k options of "author" neural methods.

(Adaku Uchendu, November 2020). The goal of the subfield of computer-generated text detection is to accomplish one or more of these goals with a high success rate.

This subfield has been active since around when natural language generation became an active subfield, with some landmark papers and books dating back to the early 2000s (particularly, Sara Laviosa's 2002 book *Corpus-based Translation Studies: Theory, Findings, Applications* comes up frequently in earlier efforts to detect machine-translated text). (D. Kurokawa, June 2009).

2 Corpora

Although this subfield has been around for over two decades, there has been a recent boom in efforts to detect text generated by transformer-based models such as GPT and GROVER, as well as some focus on Markov chains and Recurrent Neural Networks (RNNs). Thus, most of the older datasets focus on machine translation while newer ones focus on articles and social media posts created by sophisticated generative models. The following are some commonly used datasets.

2.1 Multilingual Datasets

Early papers that focus on machine translation generally pick an accepted corpus in a source language and manually translate it to the target language. Dave Carter and Diana Inkpen translated the Canadian Hansard corpus, which contained approximately 50,000 sentences, from French to English using Microsoft Bing's machine translation service. The resulting corpus is one of the most widely-cited corpora in the detection of machine translation subfield, especially for papers focusing on translation between English and French (Roee Aharoni and Goldberg, 2014), (D. Kurokawa, June 2009), (Ella Rabinovich, 2015). Papers such as (Jérémy Ferrero and Agnès, 2017) use this as a subcorpus, combining it with original machine translations of "Wikipedia articles, conference papers, [and] product reviews" to obtain an even wider corpus.

2.2 Transformers

In recent years, there has been a rise in corpora that include articles generated by GPT-2 (after training it on webtext) and real articles written by humans. For example, WebText vs. GPT-2 is a corpus consisting of articles generated by GPT-2 as well as various articles sampled from outbound Reddit links (Solaiman et al., 2019). Similarly, the Amazon Product Reviews vs. GPT-2 corpus, which contains reviews for products generated by GPT-2 and real product reviews that were scraped from Amazon's comment sections (Solaiman et al., 2019)(et al., September 2022). In addition to GPT, The Generating aRticles by Only Viewing mEtadata Records (GROVER) model is a frequently used source for such corpora, with the RealNews vs. GROVER corpus training GROVER on a set of RealNews articles and including articles the model generates, as well as some news articles that it was not trained on (Zellers et al., 2020).

2.3 Other Corpora

Although much of the recent focus has been on GPT and other transformer-based models, corpora also exist for other models and, in some cases, obfuscated generation types. For instance, the Twitter bot dataset used in various examinations of machine-generated Twitter posts relies on scraping manually-verified bot accounts on twitter and teaching a model to differentiate between those tweets and real human tweets (Tiziano Fagni, July 2020) (Julien Tourille, June 2022)(Lopez, 2021). These bot accounts often do not have publicly available generation techniques, so models trained on them can explore how general approaches to certain generation techniques can be.

3 Detection Goals

To attack the authorship attribution problem, a method of computer-generated text detection ideally is:

- Highly accurate to allow a high degree of certainty when using it.
- Data-efficient to be able to approach the theoretical 90% accuracy.
- Easy to adapt to samples from various text generative models for practicality.
- Logical to human beings so that decisions can be verified as appropriate.
- Resilient against adverserial input.

(A. Maronikolakis, June 2021) (Bin Guo, September 2019).

4 Computer-Generated Text Detection Approaches

4.1 BERT

The most common and most successful method to detect computer-generated text currently employs transformers.

One paper identified about a 50% accuracy when asking human beings to differentiate between human-generated news headlines and GPT-2 written news headlines, while the most successful model trained while writing that paper, BERT, achieved 85.7% accuracy (A. Maronikolakis, June 2021). In a similar study in which researchers attempted to detect if Arabic sentences were written by GPT-2 or humans, ARABERT (BERT with training in Arabic) was used with consistently *98.7% accuracy* (F. Harrag, December 2020). Another study used the RoBERTa (Robustly Optimized BERT Pretraining Approach) model to improve on neural-based methods by up to 10%, yielding an 87.7% accuracy (et al., September 2022). In general, BERT easily approaches or exceeds 90% accuracy, with the use of transformers dominating most recent publications of machine-generated text detection (Anton Bakhtin, June 2019). Clearly,

BERT has been performing extremely well detecting computer-generated text in the case that the entire corpus contains entirely computer-generated texts or entirely human-generated texts.

Other models for computer-generated text detection, as discussed later in this survey, are also significantly more accurate than human beings. This is not just a matter of luck – humans tend to be far worse at identifying computer-generated text than various natural language processing models. In fact, a 2020 paper by Ippolito et al. claims that "automatic detection of generated text is *easiest* when humans are fooled" (D. Ippolito, July 2020). Human accuracy is at its lowest when the excerpts are small, and at its highest with longer pieces. Still, even with multi-paragraph samples, human accuracy hovers around 70% while machine performance is consistently above 80% accuracy (Tiziano Fagni, July 2020). Interestingly, though, BERT accuracy drops drastically with the introduction of even a single human-generated word to otherwise machine-generated text, down to about 65% (D. Ippolito, July 2020).

4.2 Phrase Frequency Analysis and Fluency Detection

Another accurate method of identifying computer-generated text is phrase frequency analysis, which relies on a principle called Zipf's law. This law says that for human-written text, the second most frequent word will appear half as many times as the most frequent word, the third most frequent word will appear a third as many times as the most frequent word, and so on (Nguyen-Son, 2017). This is not the case in machine-generated writing, so this fact can be exploited to statistically differentiate between human-generated and machine-generated text. Furthermore, machines tend to generate far less complex language than human beings, with human-generated text involving more idioms, clichés, and dialectical variations (the amount of which relate to a text's *fluency*) (Hoang-Quoc Nguyen-Son, April 2019)(Nguyen-Son, 2018). By extracting frequency features using linear or logistic regression and complex phrases using ngrams and coreference resolution, human-generated text can be identified with 89% accuracy (Nguyen-Son, 2017). This high accuracy applies for studies done on languages other than English, too, with about 90% accuracy being achieved for French, Dutch, and Japanese texts (D. Kurokawa, June 2009) (Roee Aharoni and Goldberg, 2014). This accuracy is best for large texts and drops to about 77% for single-sentence queries (D. Kurokawa, June 2009).

4.3 Noise Features

Similar to the use of ngrams and frequency features, the method of using noise features relies on the predictability of human writing to distiguish between human-generated and machine-generated text. For instance, documents containing misspellings are almost exclusively human-generated, as are texts that include "spoken noise words" like "wanna" or slang like "2morrow" (Nguyen-Son, 2018). Similarly, machines tend to coin new idioms that do not otherwise exist in a language, such as when translating colloquialisms between languages – in fact, texts with such idiosyncrasies are almost exclusively machine-generated (D. Ippolito, July 2020). Combining these noise features with the above fluency features results in 80.35% accuracy and a 19.44% error rate, which was the lowest error rate as of the publication of this method (Nguyen-Son, 2018). Though improvements have been made on machine-generated text detection since then, this method laid the groundwork for impressively accurate machine-translated text detection.

5 Machine Translated Texts

5.1 "Translationese"

A straightforward way to tell when a machine has translated some text is detecting "translationese." This is a well-defined set of characteristics so common in machine-translated text that it has almost become a dialect, with the term being coined in a 2005 paper by Marco Baroni and Silvia Bernardini (Baroni and Bernardini, 2005) (Roee Aharoni and Goldberg, 2014). Though this term does not have a formal definition, humans have been tested to be able to intuitively detect it in machine-translated texts. Work done with support vector machines (SVMs) indicates that such a machine dialect can be identified with with 86.7% accuracy. The studies suggest that the distribution of morphosyntactic units, pronouns, prepositions, and adverbs tend to connote the presence of translationese, and thus machine-generated text (Baroni and Bernardini, 2005) (Ella Rabinovich, 2015).

5.2 Statistical Machine Translation Detection

The result of such prominent features in machine-translated text and different, similarly prominent features in human-generated text results in effective models in statistical detection of machine translation. By extracting features such as "phrase salads," wherein there are obviously mistranslated idioms in text; correctly used idioms and figures of speech; and evidence of translationese, researchers are able to detect machine translation with 95.8% accuracy for short texts and 80.6% accuracy for webpages (all done in English) (Arase and Zhou, 2013)(D. Ippolito, July 2020)(Nguyen-Son, 2017). Training models in parts of speech tagging in order to detecting grammatical and tense-based errors, as well as using bigrams and other n-grams, further raises this accuracy for webpages to 85% (D. Kurokawa, June 2009).

5.3 Back-translation and Round-Trip Translation

Back-translation refers to translating a piece of text from the language it is in to the target language of the user (Michael Jones, January 2014). Round-trip translation goes a step further, translating a body of work into a different language and then translating it back to its original language, resulting in text with a very similar essence but few direct similarities in diction (Hoang-Quoc Nguyen-Son and Kiyomoto, 2019). Although this is very easy to detect when done with the same machine translator, with the resulting text matching almost exactly, it can be hard to detect when the two translation steps are done using two different translators. Back-translation and round-trip translations primarily have unethical uses, with the main one being plagiarism. These methods subvert common plagiarism-detection methods such as similarity checking and human examination (Michael Jones, January 2014). The increase in machine translation availability over the past decade is making the prevelance of these actions in academic and professional settings an ever-growing problem that, luckily, now has some fairly accurate detection methods (H. Nguyen-Son, April 2021).

Methods to detect back- and round-trip translation were initially created as early as 2008 with with the creation of multilingual databases to detect synonyms and perform cross-language similarity checking (Hoang-Quoc Nguyen-Son and Kiyomoto, 2019). Such methods are extremely bloated, though, and would require incredible amounts of information to check for plagiarism between just two language, let alone all possible languages. Even worse, they have about 60% accuracy – clearly not worth the amount of resources it would use up (Michael Jones, January 2014).

More sophisticated techniques from recent years to detect such ploys have had more success. Though still working with on text similarity with round-trip translation (TSRT), a 2021 method relies on Convolutional Neural Networks (CNNs) trained on original and translated texts in various languages. This method has a 86.9% average accuracy over the course of the study, increasing to about 97% as text lengths get longer (Hoang-Quoc Nguyen-Son and Kiyomoto, 2019). Word embeddings can also be used in this context, having additional particular accuracy in determining if a document was manually translated into another language with about 85% accuracy (Anton Bakhtin, June 2019)(Jérémy Ferrero and Agnès, 2017). Similar methods could further determine which translator was used to complete a certain translator, as well as which language the text was written in prior to translation, with 93.3% and 85.6% accuracy respectively.

One method from 2019 was able to use back-translation for the purpose of determining if a given text is machine-generated, using BLEU scores to calculate the similarity between an input text and its round-trip translation. The study claims that the use of a classifier at this step allows this determination, achieving 75.0% accuracy (Michael Jones, January 2014).

6 Limitations and Future Work

Although impressively accurate already, detection of machine-based text still requires much work. In particular, accuracy of methods decreases substantially when human-generated text is added to computer-generated text (D. Ippolito, July 2020). Additionally, many existing methods fare poorly in the face of newer concepts such as Large-Scale Language Models (LLMs), which are capable of generating very long, logically consistent documents (L.R. Varshney, February 2020). Furthermore, technology such as ChatGPT and Google Translate are readily available to the public and built into many people's daily lives at this point, while technology to counter misinformation generated by such tools is not yet widely available

(Bin Guo, September 2019). Though this subfield has evolved very quickly over the last five years, it must continue to develop at breakneck speeds if it is to effectively combat malicious campaigns from generative language tools.

7 Citations

References

M. Stevenson A. Maronikolakis, H. Schutze. June 2021. Identifying automatically generated headlines using transformers. *Association for Computational Linguistics*, Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda(2021.nlp4if-1.1):1–6.

Kai Shu Dongwon Lee Adaku Uchendu, Thai Le. November 2020. Authorship attribution for neural text generation. *Association for Computational Linguistics*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)(2020.emnlp-main.673):189–196.

Myle Ott Yuntian Deng Marc'Aurelio Ranzato Arthur Szlam Anton Bakhtin, Sam Gross. June 2019. Real or fake? learning to discriminate machine from human generated text. *arXiv:1906.03351 [cs.LG]*.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. *Association for Computational Linguistics*, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)(P13-1157):1597–1607.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

Lina Yao Yunji Liang Zhiwen Yu Bin Guo, Yasan Ding. September 2019. The future of misinformation detection: New perspectives and trends. *arXiv:1909.03654 [cs.SI]*.

Chris Callison-Burch Douglas Eck D. Ippolito, Daniel Duckworth. July 2020. Automatic detection of generated text is easiest when humans are fooled. *Association for Computational Linguistics*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics(2020.acl-main.164):1808–1822.

P. Isabelle D. Kurokawa, C. Goutte. June 2009. Automatic detection of translated text and its impact on machine translation. *Association for Computational Linguistics*, (2009.mtsummit-papers).

Shuly Wintner Ella Rabinovich. 2015. Unsupervised identification of translationese. *Association for Computational Linguistics*, Transactions of the Association for Computational Linguistics, vol. 3(Q15-1030):419–432.

L. Kushnareva1 et al. September 2022. Artificial text detection via examining the topology of attention maps. *Association for Computational Linguistics*, (2021.emnlp-main.50).

K. Darwish-A. Abdelali F. Harrag, M. Debbah. December 2020. Bert transformer model for detecting arabic gpt2 auto-generated tweets. *Association for Computational Linguistics*, Proceedings of the Fifth Arabic Natural Language Processing Workshop(2020.wanlp-1.19):207–214.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation.

S. Hidano I. Gupta S. Kiyomoto H. Nguyen-Son, T.P. Thao. April 2021. Machine translated text detection through text similarity with round-trip translation. *Association for Computational Linguistics*.

Seira Hidano Hoang-Quoc Nguyen-Son, Thao Tran Phuong and Shinsaku Kiyomoto. 2019. Detecting machine-translated text using back translation. *Association for Computational Linguistics*, Proceedings of the 12th International Conference on Natural Language Generation(W19-8626):189–197.

Seira Hidano Shinsaku Kiyomoto Hoang-Quoc Nguyen-Son, Tran Phuong Thao. April 2019. Detecting machine-translated paragraphs by matching similar words. *arXiv:1904.10641 [cs.CL]*.

Adrian Popescu Julien Tourille, Babacar Sow. June 2022. Automatic detection of bot-generated tweets.

Didier Schwab Jérémy Ferrero, Laurent Besacier and Frédéric Agnès. 2017. Using word embedding for cross-language plagiarism detection. *Association for Computational Linguistics*, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers(E17-2066):415–421.

Gallemore C. Lopez, C.E. 2021. An augmented multilingual twitter dataset for studying the covid-19 infodemic.

R. Socher L.R. Varshney, N.S. Keskar. February 2020. Limits of detecting text generated by large-scale language models. *arXiv:2002.03438 [cs.CL]*.

Lynnaire Sheridan Michael Jones. January 2014. Back translation: an emerging sophisticated cyber strategy to subvert advances in 'digital age' plagiarism detection and prevention. *University of Wollongong*.

Echizen I. Nguyen-Son, HQ. 2018. Detecting computer-generated text using fluency and noise features. *Hasida, K., Pa, W. (eds) Computational Linguistics. PACLING 2017. Communications in Computer and Information Science, vol 781.*

T. Tieu N-D H. Nguyen H Yamagishi J Echizen I Nguyen-Son, H-Q. 2017. Identifying computer-generated text using statistical analysis. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*.

Moshe Koppel Roee Aharoni and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. *Association for Computational Linguistics*, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)(P14-2048):289–295.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release strategies and the social impacts of language models.

Margherita Gambini Antonio Martella Maurizio Tesconi Tiziano Fagni, Fabrizio Falchi. July 2020. Tweepfake: about detecting deepfake tweets. *arXiv:2008.00036 [cs.CL]*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news.