

Detecting Tortured Phrases in Scientific Paper

LAY Puthineath



Table of contents



01. Introduction

02. Problems

03. Solutions



01

Introduction





Improper peer review of scientific paper?



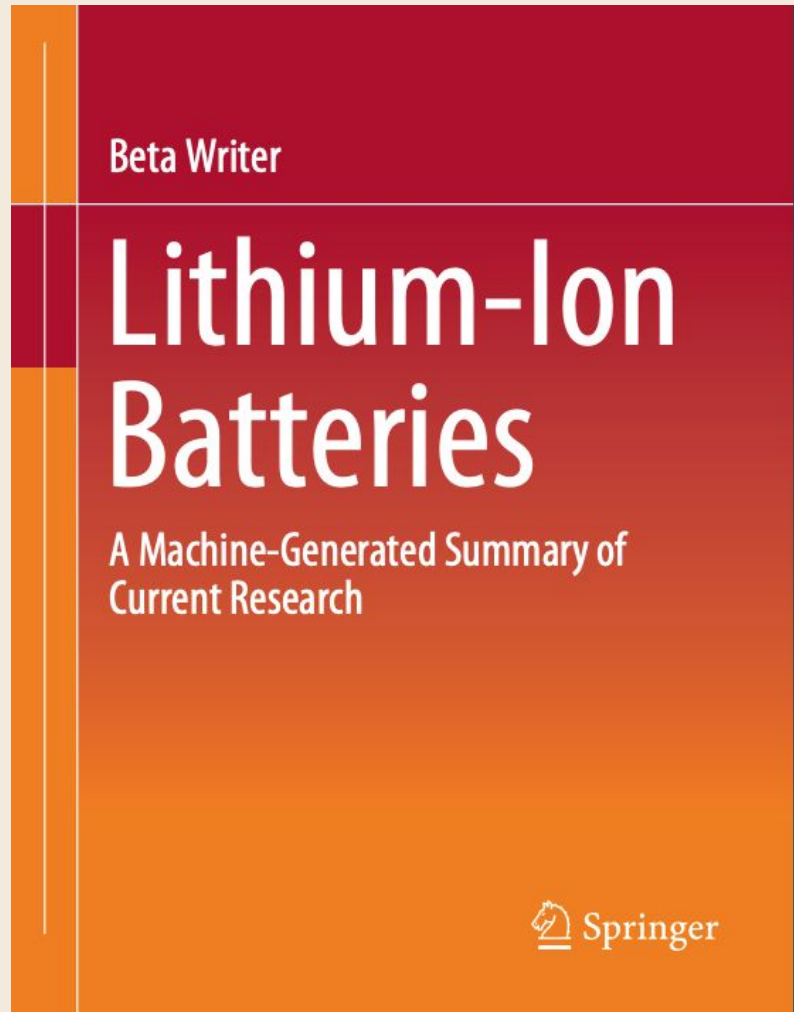
“They create journals with names like the **American Journal of Medical and Dental Sciences** or the **European Journal of Chemistry** to imitate—and in some cases, literally clone—those of **Western** academic publishers. But the locations revealed by **IP addresses** and **bank invoices** are continents away”

Bohannon, J. (2013)





New books are created with the inputs of thousand articles. **Ex:** The 278-page book is the first from Springer Nature that author is the machine, not human.



How nonsensical papers are produced?



Human-written texts



Machine-generated texts

Mathgen

Randomly generated mathematics research papers!

[About](#) | [Buy a book](#) | [Get the code](#) | [Blog](#) | [SCIgen](#)

Produce your own math paper, full of research-level, professionally formatted nonsense! Just enter your name and those of up to 3 "co-authors":

Author 1: A. Lastname or: generic name famous name
Author 2: or: generic name famous name
Author 3: or: generic name famous name
Author 4: or: generic name famous name

SCIgen - An Automatic CS Paper Generator

[About](#) [Generate](#) [Examples](#) [Talks](#) [Code](#) [Donations](#) [Related](#) [People](#) [Blog](#)

About

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. It uses a hand-written **context-free grammar** to form all elements of the papers. Our aim here is to maximize amusement, rather than coherence.

One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. A prime example, which you may recognize from spam in your inbox, is SCI/IIIS and its dozens of co-located conferences (check out the very broad conference description on the [WMSCI 2005](#) website). There's also a list of [known bogus conferences](#). Using SCIgen to generate submissions for conferences like this gives us pleasure to no end. In fact, one of our papers was accepted to SCI 2005! See [Examples](#) for more details.

We went to WMSCI 2005. Check out the [talks and video](#). You can find more details in our [blog](#).

Also, check out our 10th anniversary celebration project: [SCIpher!](#)

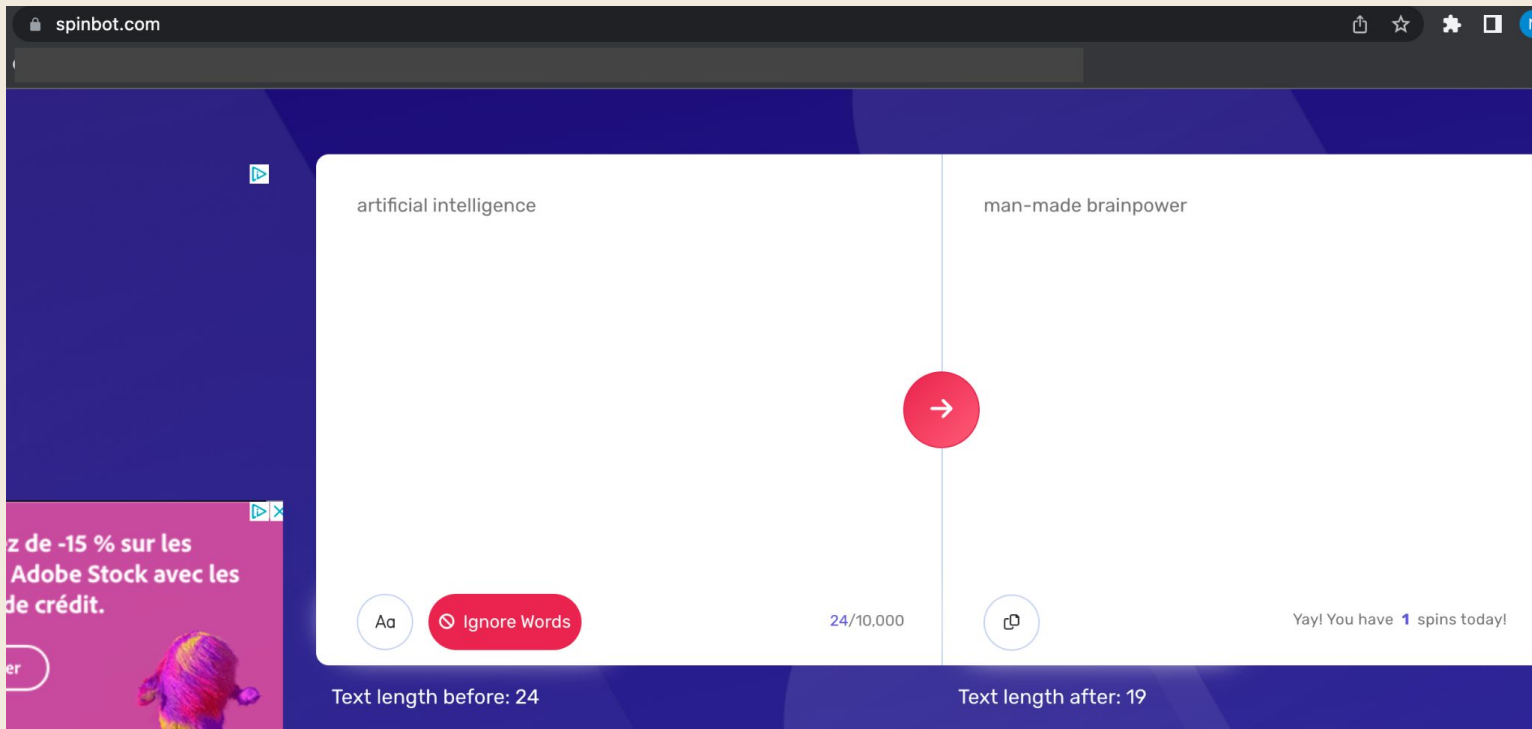
Generate a Random Paper

Want to generate a random CS paper of your own? Type in some optional author names below, and click "Generate".

Author 1:
Author 2:
Author 3:
Author 4:
Author 5:

Generated mathematics research papers websites

Spinbot: paraphrasing tool



The screenshot shows the Spinbot website interface. The browser address bar displays "spinbot.com". The main content area is split into two columns. The left column contains the text "artificial intelligence" and "Text length before: 24". The right column contains the text "man-made brainpower" and "Text length after: 19". A red circular button with a white right-pointing arrow is positioned between the two columns. At the bottom of the interface, there are several controls: a font style icon "Aa", a red "Ignore Words" button, a character count "24/10,000", a copy icon, and a notification "Yay! You have 1 spins today!". A pink promotional banner is visible on the left side of the interface.

Artificial Intelligence -> man-made brainpower



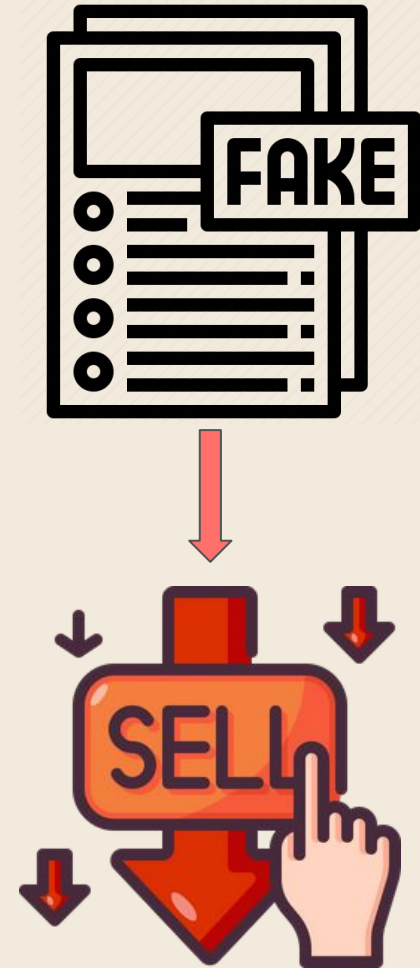
02

Problems



“As a result, meaningless randomly generated scientific papers end up being served and sometimes sold by various publishers with a prevalence estimated to 4.29 papers every one million papers”.

(Cabanac & Labbé, in press; Van Noorden, 2021)



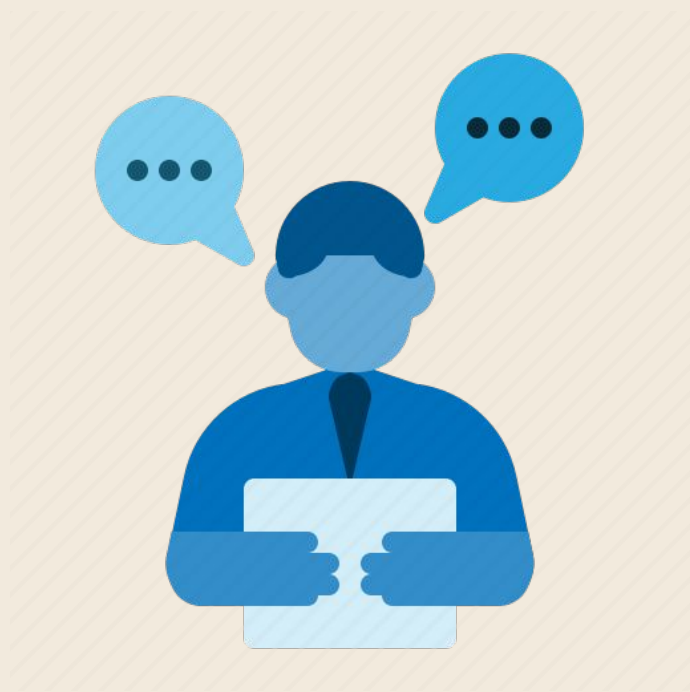


Tortured Phrases:



unexpected weird phrases instead of the established ones, such as “counterfeit consciousness” or “man-made brain power” instead of “artificial intelligence”.

Only human can detect the **tortured**
phrases currently





03

Solutions





Objective



Based on the current machine learning technology and language model, my study is to detect the new kind of the tortured phrases in the sentences automatically.



Ex: It is commonly acknowledged that FDI is one of the essential wellsprings of capital inflow and driving components of financial development in many **creating nations**.



creating nations is the tortured phrase.

developing countries is the expected phrase.



Current study



Investigating on various experiment to differentiate the **tortured phrases** and **expected phrases** based on classification techniques.



References

- Beta Writer. (2019). *Lithium-Ion batteries: A machine-generated summary of current research*. Springer. (Machine-generated by Beta Writer 0.7 software developed at Goethe University Frankfurt) doi: 10.1007/978-3-030-16800-1
- Bohannon, J. (2013). Who's afraid of peer review? [News]. *Science*, 342(6154), 60–65. doi: 10.1126/science.342.6154.60
- Charles Day. Here come the robot authors. *Physics Today*, 72(6):8–8, 2019.
- Cabanac, G., & Labbé, C. (in press). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.24495
- Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. Tortured phrases: A dubious writing style emerging in science. evidence of critical issues affecting established journals. *CoRR*, abs/2107.06751, 2021.

Thanks

Do you have any
questions?

CREDITS: This presentation template was created by **Slidesgo**,
including icons by **Flaticon**, infographics & images by **Freepik**

