Découverte d'acronymes torturés dans des publications scientifiques

Alexandre Clausse

Découverte d'acronymes torturés dans des publications scientifiques

Alexandre CLAUSSE

Université Toulouse III – Paul Sabatier
Institut de recherche en informatique de Toulouse (IRIT UMR 5505 CNRS)
118 route de Narbonne
31062 Toulouse cedex 9

alexandre.clausse@univ-tlse3.fr

RÉSUMÉ : *Dans un contexte de course à la publication, du contenu plagié est régulièrement publié, amenant à une pollution croissante de la littérature scientifique. Une telle fraude peut être caractérisée par l'utilisation d'expressions torturées. Des solutions ont été développées afin de contribuer à la détection et au signalement de tels contenus. D'une part, elles reposent sur des méthodes et ressources de nature hétérogène, avec des biais qui leur sont propre. D'autre part, elles nécessitent la collaboration d'experts permettant l'alimentation d'une liste d'expressions connues. Ainsi, nous proposons une approche peu coûteuse en ressources et indépendante de tout domaine, reposant sur la détection d'acronymes torturés, étant visuellement facile à mettre en œuvre. Afin de détecter la présence de l'ensemble de ces expressions dans une publication donnée, nous mettons à disposition un jeu de données de publications torturées, un algorithme d'extraction et de classification d'acronymes, ainsi qu'une méthode permettant d'évaluer cette ligne de base. Les résultats obtenus sont biaisés par l'utilisation du jeu de données de développement, annoté par une seule personne, lors de l'évaluation de la solution proposée. Nos futures recherches seront focalisées sur l'élaboration de méthodes permettant la détection de formes particulières d'expressions telles que les hallucinations et les termes polysémiques, toujours dans une optique de faciliter la détection d'expressions torturées.*

MOTS-CLÉS : *Extraction d'information, détection d'anomalie, plagiat, intégrité scientifique.*

ENCADREMENT : *Guillaume Cabanac, Pascal Cuxac, Cyril Labbé.*

1. Contexte

Les politiques de recherche publique exercent une pression constante sur les chercheurs, en les incitant à publier le plus régulièrement possible dans des revues réputées, afin d'obtenir les meilleures performances dans les classements internationaux. Cette situation, communément appelée « publier ou périr », amène un petit nombre de personnes peu scrupuleuses à manquer de considération pour la rigueur nécessaire aux travaux de recherche scientifique, en ayant recours à la fabrication, à la falsification et au plagiat. Ces problèmes ont été décrits dans un ouvrage édité par Biagioli *et al.* (2020). Le plagiat peut être déguisé par l'utilisation d'expressions torturées, définies par Cabanac *et al.* (2021) comme étant le remplacement de termes scientifiques établis par des synonymes, les vidant ainsi de toute signification, principalement en utilisant des outils de paraphrase. Par exemple, le terme informatique « *machine learning* » peut être paraphrasé en « *computer mastering* ». De plus, un texte contenant des expressions torturées peut échapper à la vigilance d'un comité de relecture, et être publié tel quel, amenant à une pollution croissante de la littérature scientifique. Cela met aussi en doute leur probité quant à la tenue des expériences et la rédaction sincère des résultats. Pour contrer un tel problème, il est envisageable de permettre aux différents acteurs de l'édition savante de vérifier la conformité d'un contenu, notamment vis-à-vis du respect des valeurs et principes d'honnêteté et de rigueur attendu de la part des chercheurs. Une solution consiste à développer une brique logicielle permettant d'identifier en amont de la publication les articles contenant de telles expressions, par leur détection automatique, dans le but d'y apporter une attention accrue.

2. État de l'art

En 2021, le *Problematic Paper Screener* (PPS)[1] a été développé pour permettre le recensement de publications scientifiques frauduleuses ou suspectes, incluant les contenus plagiés par l'utilisation d'expressions torturées. Ce site facilite l'évaluation post-publication en identifiant des articles problématiques et en préparant des rapports d'évaluation que les utilisateurs peuvent publier sur PubPeer[2], une plateforme de réévaluation collaborative d'articles scientifiques. Ainsi, plus de 5 000 expressions torturées distinctes ont été recensées, et plus de 13 000 articles contenant de telles expressions ont été listés sur le PPS. Dans l'optique de détecter du contenu plagié, des travaux ont constitué des corpora de documents web, journalistiques et scientifiques, pour expérimenter plusieurs approches utilisant des modèles de langue et d'apprentissage machine. Gehrmann *et al.* (2019) ont proposé une méthode de détection de textes générés par des modèles de langue, axée sur des caractéristiques syntaxiques et distributionnelles. Les résultats obtenus présentent un biais lié à l'utilisation de données similaires. Wahle *et al.* (2022) ont cherché des façons d'identifier le plagiat par paraphrasage en constituant un jeu de données composé de paragraphes paraphrasés, et

1. https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener
2. https://www.pubpeer.com

en comparant leurs résultats de classification avec ceux de deux détecteurs de plagiat. Leurs résultats sont biaisés par l'utilisation d'un jeu d'entraînement sur les mêmes poids lors de l'apprentissage par transfert de leurs modèles d'apprentissage machine. Lay *et al.* (2022) ont complété l'étude précédente par l'apport d'un ensemble de cinq-grammes (suite de cinq mots), comprenant ou non des expressions torturées. Leurs résultats varient d'un modèle et d'une fonction d'agrégation à l'autre. De plus, les résultats de classification sont biaisés par l'utilisation d'un jeu de données déséquilibré, dont certaines sont dupliquées. Martel *et al.* (2023) ont également réutilisé ce corpus afin de réaligner une centaine de paires de phrases et compléter l'étude de Wahle *et al.* (2022), leurs résultats présentent des biais similaires, en plus de comporter des caractéristiques peu exploitables car extraites depuis des modèles de plongement lexical non contextuel, posant un problème de polysémie.

3. Problématique

Étant données les limites soulignées dans la section 2, il faudrait arriver à extraire d'autres caractéristiques permettant la détection d'expressions torturées. De plus, les méthodes actuelles de détection se font par lecture humaine et la collecte participative de nouvelles expressions torturées (notamment via le PPS). Il faudrait donc automatiser cette tâche et compléter ces méthodes afin de se prémunir d'autres formes de plagiat.

4. Contribution

Très récemment, nous avons exploré l'extraction d'acronymes torturés avec un algorithme de correspondance (notamment en comparant les initiales), développé en utilisant un corpus vérité-terrain de 75 articles scientifiques manuellement annotés (Clausse *et al.*, 2024). Les résultats obtenus sont biaisés car l'algorithme a été évalué sur l'ensemble de données de développement, et notre approche ne peut pas détecter les formes hallucinées (par exemple, le « *Bolster Vector Machine (BVM)* » où l'acronyme correspond à sa forme développée). Les prochaines étapes consistent donc à améliorer la qualité du corpus vérité-terrain par l'annotation de celui-ci (par au moins une autre personne) et le calcul d'un accord inter-annotateur.

5. Recherches futures

Dans la mesure où nous travaillons sur des données textuelles en anglais et dans un nombre restreint de domaines, il est nécessaire de disposer d'ensembles de données axés sur des articles scientifiques torturés provenant de plusieurs domaines, générés par différents outils de paraphrase et dans différentes langues. Des échantillons synthétiques pourraient être paraphrasés à l'aide d'auto-encodeurs ou générés à l'aide de modèles auto-régressifs (les expressions torturées recensées par le PPS pourraient

être utilisées), et pourraient être comparés à des échantillons existants. Les domaines pourraient être pris en compte à l'aide de la modélisation thématique. Dans le but de détecter les expressions torturées en utilisant des ressources informatiques limitées, il serait intéressant de prendre en compte plusieurs caractéristiques pour réaliser des algorithmes de correspondance, avec la difficulté de prendre en compte les différentes formes de ces expressions (par exemple les acronymes ou encore la spécificité d'un domaine). Il serait aussi intéressant de développer une méthode permettant d'identifier les sources de plagiat. Pour cela, l'utilisation de modèles de langues permettrait de détecter du contenu suspect, associés à des dictionnaires, pour retrouver le contenu original en fonction d'un seuil de score de similarité.

6. Bibliographie

Biagioli M., Lippman A. (éd.), *Gaming the Metrics: Misconduct and Manipulation in Academic Research*, The MIT Press, 2020. DOI: https://doi.org/10.7551/mitpress/11087.001.0001.

Cabanac G., Labbé C., Magazinov A., "Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals", 2021. Preprint arXiv. DOI: https://doi.org/10.48550/arXiv.2107.06751.

Clausse A., Cabanac G., Cuxac P., Labbé C., "Extraction d'acronymes torturés dans la littérature scientifique", *in* P. Cuxac, C. Lopez (éd.), *Atelier TextMine de la conférence Extraction et Gestion des Connaissances (EGC) de 2024*, Dijon, France, p. 27–37, 2024. URL : https://hal.science/hal-04426448.

Gehrmann S., Strobelt H., Rush A., "GLTR: Statistical Detection and Visualization of Generated Text", *in* M. R. Costa-jussà, E. Alfonseca (éd.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, p. 111–116, 2019. DOI: https://doi.org/10.18653/v1/P19-3019.

Lay P., Lentschat M., Labbe C., "Investigating the detection of Tortured Phrases in Scientific Literature", *in* A. Cohan, G. Feigenblat, D. Freitag, T. Ghosal, D. Herrmannova, P. Knoth, K. Lo, P. Mayr, M. Shmueli-Scheuer, A. de Waard, L. L. Wang (éd.), *Proceedings of the Third Workshop on Scholarly Document Processing*, Association for Computational Linguistics, p. 32–36, 2022. URL: https://aclanthology.org/2022.sdp-1.4.

Martel E., Lentschat M., Labbe C., "Detection of Tortured Phrases in Scientific Literature", *in* T. Ghosal, F. Grezes, T. Allen, K. Lockhart, A. Accomazzi, S. Blanco-Cuaresma (éd.), *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, Association for Computational Linguistics, p. 43–48, 2023. DOI: https://doi.org/10.18653/v1/2023.wiesp-1.6.

Wahle J. P., Ruas T., Foltýnek T., Meuschke N., Gipp B., "Identifying Machine-Paraphrased Plagiarism", *in* M. Smits (éd.), *Information for a Better World: Shaping the Global Future*, Springer, p. 393–413, 2022. DOI: https://doi.org/10.1007/978-3-030-96957-8_34.