Extraction d'acronymes torturés dans la littérature scientifique

Alexandre Clausse, Guillaume Cabanac, Pascal Cuxac, Cyril Labbé

Extraction d'acronymes torturés dans la littérature scientifique

Alexandre Clausse*, Guillaume Cabanac*,**, Pascal Cuxac*** et Cyril Labbé****

*Université Toulouse III – Paul Sabatier, IRIT UMR 5505 CNRS, Toulouse, France
{alexandre.clausse, guillaume.cabanac}@irit.fr

**Institut Universitaire de France (IUF), Paris, France

***INIST – CNRS, UAR76, Vandœuvre-lès-Nancy, France
pascal.cuxac@inist.fr

****Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France
cyril.labbe@univ-grenoble-alpes.fr

Résumé. Les politiques publiques poussent les chercheurs à publier le plus régulièrement possible des articles scientifiques dans des revues réputées. Cette pression amène une minorité de personnes peu scrupuleuses à frauder, en utilisant notamment des phrases torturées afin de déguiser des plagiats. Il arrive que ce type de contenu ne soit pas détecté lors de l'évaluation par les pairs, amenant à sa publication. Dans l'optique de dépolluer la littérature scientifique, nous proposons une tâche d'extraction d'acronymes torturés permettant la détection de publications frauduleuses. Un benchmark est effectué grâce à un corpus de 75 articles scientifiques torturés et en open access, avec une chaîne de traitements permettant l'extraction et de classification d'acronymes. La tâche d'extraction d'acronymes torturés obtient une F-mesure de 0,74, qui pourra être améliorée par l'enrichissement du corpus. Nous avons constitué une ligne de référence *(baseline)* à laquelle toute personne désireuse d'améliorer la performance de cette tâche pourra se référer.

Introduction

Les politiques publiques actuelles exercent une constante pression sur les chercheurs, en les poussant à publier le plus régulièrement possible dans des revues réputées (à fort facteur d'impact), afin d'obtenir de meilleures performances dans les classements internationaux. Communément appelée « publier ou périr », cette situation est susceptible d'amener à un manque de considération pour la rigueur impliquée dans les travaux de recherche scientifique par une minorité de personnes peu scrupuleuses, ayant recours à de la fabrication, de la falsification et au plagiat. Ces problèmes ont été décrits dans un ouvrage édité par (Biagioli et Lippman, 2020). Le plagiat peut être déguisé par l'utilisation de phrases torturées, définies par (Cabanac et al., 2021), correspondant à la déformation d'un phraséoterme (terme scientifique considéré comme établi dans une discipline associée), par la génération de synonymes vidant ce dernier

de toute signification, principalement en utilisant des outils de paraphrase (*spinners*) tels que SpinBot [1]. Par exemple, le terme informatique « *artificial intelligence* » peut être transformé en « *man-made consciousness* », ce qui n'a aucune signification dans un tel domaine métier. De plus, un texte comprenant des phrases torturées peut échapper à la vigilance d'un comité de relecture, amenant à sa publication, et par conséquent à la pollution de la littérature scientifique. Cela met aussi en doute leur probité quant à la tenue des expériences et la rédaction sincère des résultats.

Étant donné ce contexte, le *Problematic Paper Screener* [2] (PPS) a été développé par (Cabanac et al., 2022) afin de permettre la réévaluation collaborative, par le biais de PubPeer [3] (créé et administré par (Barbour et Stell, 2020)), d'articles scientifiques signalés comme étant suspects. Parmi un ensemble de détecteurs (permettant par exemple de détecter des articles torturés ou générés par SciGen [4]), 13 000 articles ont été considérés comme contenant suffisamment de phrases torturées pour que leur rigueur scientifique soit remise en cause. La détection de tels contenus a été explorée par (Lay et al., 2022) par la construction d'un corpus de 2 772 paragraphes contenant ou non des phrases torturées. Ils ont utilisé des algorithmes de classification binaire (forêt d'arbres décisionnels, perceptron, transformeur) sur les 5-grammes et paragraphes. Leur meilleur modèle a obtenu une F-mesure de 0,99 (classe « torturée ») et 0,92 (classe « non torturée »). Cependant cette approche est limitée par la possible présence des mêmes phrases torturées dans les données d'entraînement et de test, car celles-ci ont été séparées de façon aléatoire. (Kashnitsky et al., 2022) ont proposé une tâche partagée avec un corpus de plus de 26 000 documents extraits depuis Scopus, afin de détecter différentes formes de textes générés (par exemple par résumé automatique ou par l'utilisation de *spinners*). Ils ont utilisé la régression logistique ainsi qu'un modèle SciBERT ((Beltagy et al., 2019)) ajusté comme ligne de base, en ayant respectivement obtenu une F-mesure de 0,82 et 0,98. (Becker et al., 2023) ont étendu cette exploration par un corpus totalisant plus de 140 000 paires de phrases et paragraphes paraphrasés par un humain ou une machine. Ils ont utilisé des modèles de plongement lexical tels que GloVe ((Pennington et al., 2014)) et FastText ((Joulin et al., 2016)), ainsi que des modèles de langage tels que BERT ((Devlin et al., 2019)) et T5 ((Raffel et al., 2020)), leur meilleur modèle a obtenu une F-mesure de 0,95.

Cet article présente la problématique de la thèse que le premier auteur vient de débuter, par une approche orientée document et focalisée sur les acronymes torturés (c'est-à-dire des phraséotermes sous forme d'acronymes qui ont été torturés, par exemple « *fake neural system (ANN)* » est une version torturée de « *artificial neural network* »). Nous proposons un algorithme auquel est associé un benchmark, par l'utilisation d'un corpus composé d'articles scientifiques en accès libre. Sont adjoints une chaîne de traitements permettant l'extraction et la classification d'acronymes, ainsi que des métriques d'évaluation des résultats obtenus.

Constitution d'une collection de test

À partir du PPS (détecteur « *tortured* »), nous avons collecté un corpus de 75 articles scientifiques en accès libre, dans le domaine de l'ingénierie (dont les disciplines sont représentés

1. https://spinbot.com
2. https://irit.fr/~Guillaume.Cabanac/problematic-paper-screener
3. https://pubpeer.com
4. https://pdos.csail.mit.edu/archive/scigen/

en Figure 1), publiés entre 2015 et 2023 (dont la distribution des années de publication est représentée en Figure 2). Nous avons effectué une recherche par mot-clé dans le champs « *venue* », c'est-à-dire le titre de la revue ou conférence dans laquelle les articles collectés ont été publiés, en utilisant le nom de 10 disciplines en anglais (par exemple, les articles du domaine de la physique ont été collectés en effectuant une recherche avec le terme « *physics* » présent dans le titre de la revue associée, en vérifiant qu'ils sont bien en accès libre). Malgré cela, notre corpus comprend 37 articles pluridisciplinaires, dont la discipline commune est l'informatique. Puis nous avons effectué une analyse exploratoire des données afin d'en extraire les principales caractéristiques.

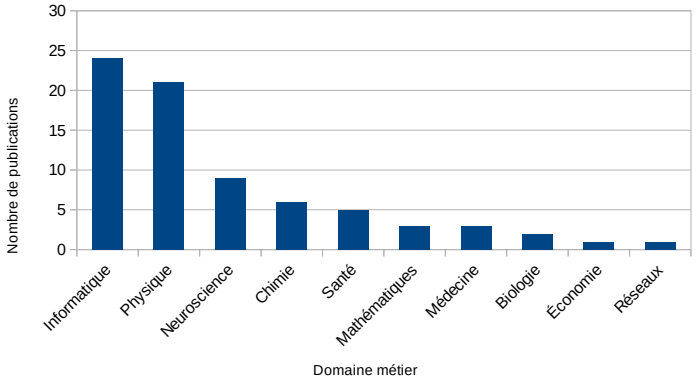

FIG. 1 – *Distribution des domaines métiers dans le corpus. Plus de la moitié de celui-ci contient des articles des domaines de l'informatique (24 articles soit 32 % du corpus) et de la physique (21 articles soit 28 % du corpus), les autres domaines étant peu représentés (entre 1 et 9 articles soit entre 1,3 et 12 % du corpus).*

Le premier auteur a manuellement extrait et annoté 995 acronymes distincts, établissant ainsi un *silver standard*, accessible publiquement (voir la section disponibilité des données). Nous avons récupéré l'ensemble des acronymes torturés (366 acronymes soit 36,8 % du corpus), plusieurs acronymes comportant des fautes de frappe (par exemple « *Tru_-Positive (TP)* ») ont été étiquetés comme suspects (46 acronymes soit 4,6 % du corpus), les autres acronymes ont été annotés comme non torturés (indiqués comme *genuine*). Puis nous avons extrait un ensemble de caractéristiques (décrites dans le Tableau 1 et dont la distribution est décrite en Figure 3) à prendre en compte dans le développement d'un algorithme d'extraction et de classification.

Parmi ces caractéristiques, nous avons remarqué quelques subtilités. La différence entre le nombre d'initiales d'un phraséoterme et son acronyme associé est principalement due à la présence de mots de liaison. Dans certains cas, un mot supplémentaire est présent entre l'acronyme et sa forme développée. Certains acronymes sont déclinés dans leur forme plurielle (par exemple « *CNN* » et « *CNNs* »). D'autres acronymes contiennent des caractères spéciaux tels que des parenthèses (cela concerne principalement les composés chimiques tels que « *poly(ethylene glycol) (PEG)* »), des points d'interrogation (cela concerne les erreurs d'encodage des lettres grecques, telles que « *Tumor Necrosis Factor Alpha (TNF- ?)* », liées à la

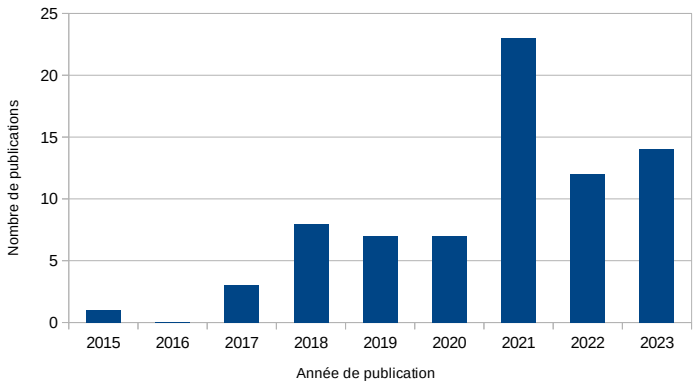

FIG. 2 – *Distribution des années de publication des articles du corpus. Nous pouvons observer une augmentation significative de publications contenant des phrases torturées à partir de 2018, avec un pic en 2021. Ce pic correspond à un grand nombre d'articles publiés en 2021 en accès libre ayant pu être récupérés depuis une même source.*

conversion des fichiers PDF bruts en texte, voir la Figure 4), des tirets et des points. Une quantité négligeable d'acronymes est rédigée dans une forme alternative (par exemple « *(PLSR) Partial Least Squares Regression* » où l'acronyme en parenthèses est positionné avant sa forme développée, et « *CNN (Convolutional Neural Network)* » où c'est la forme développée qui se trouve entre parenthèses au lieu de l'acronyme). Nous évaluons trois tâches distinctes : extraction d'acronymes, classification d'acronymes, extraction d'acronymes torturés. Pour cela, nous calculons les scores de rappel, précision, F-mesure et — dans la mesure du possible — le coefficient de corrélation de Matthews (*MCC*) ainsi que l'aire sous la courbe de la fonction d'efficacité du récepteur (*ROC AUC*). Ces métriques d'évaluation sont calculées différemment d'une tâche à l'autre, de par la nature de celles-ci. En effet, nous comptons différemment les vrai-positifs (VP), vrai-négatifs (VN), faux-positifs (FP) et faux-négatifs (FN) pour la tâche d'extraction (recherche d'information) et la tâche de classification.

Concernant l'évaluation de la tâche d'extraction, les acronymes correctement extraits depuis le *silver* sont considérés comme vrai-positifs (VP), les acronymes extraits qui n'apparaissent pas dans le *silver* sont considérés comme faux-positifs (FP), et les acronymes qui n'ont pas été extraits depuis le *silver* sont considérés comme faux-négatifs (FN). Nous ne calculons pas de vrai-négatifs (VN) dans la mesure où ils seraient définis comme des acronymes absents du *silver* et non extraits, or nous n'avons aucun moyen d'en obtenir. Concernant l'évaluation de la tâche de classification, les acronymes torturés sont considérés comme positifs, et ceux non torturés comme négatifs, les vrai- et faux-positifs / négatifs sont définis en comparant l'étiquetage des acronymes du *silver* et de ceux classifiés. Enfin, concernant l'évaluation de la tâche d'extraction des acronymes torturés, les VP correspondent aux acronymes torturés du *silver* étiquetés en tant que tel par l'algorithme. Les FP correspondent à tous les acronymes non torturés du *silver* qui n'ont pas été extraits ou qui ont été étiquetés comme torturés mais aussi aux acronymes étiquetés comme torturés mais qui n'apparaissent pas dans le *silver*. Les FN correspondent à tous les acronymes torturés du *silver* qui n'ont pas été extraits ou qui ont été étiquetés comme non torturés mais aussi aux acronymes étiquetés comme non torturés mais

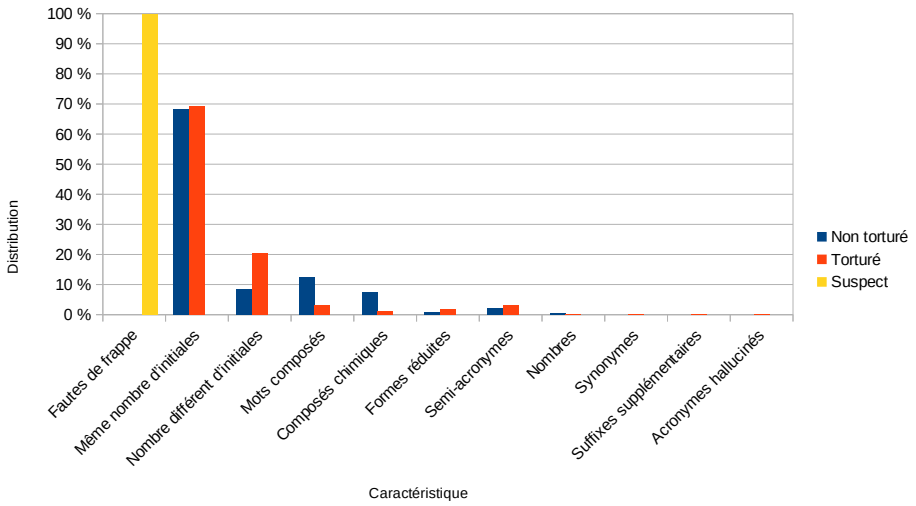

FIG. 3 – *Distribution des caractéristiques des acronymes dans chaque classe associée. La distribution des acronymes torturés comportant un nombre différent d'initiales est plus importante que pour les acronymes non torturés, cela peut être expliqué par la génération de synonymes composés de plusieurs mots. Au contraire, il y a moins d'acronymes torturés comprenant des mots composés et des composés chimiques, car il est plus difficile de trouver des synonymes à ceux-ci.*

qui n'apparaissent pas dans le *silver*. Nous ne calculons pas de VN dans la mesure où ils seraient définis comme des acronymes non torturés dans le *silver* et étiquetés comme tel, ce qui est en dehors du contexte de cette tâche.

Il est important de préciser que les acronymes considérés comme suspects dans le *silver* ont été étiquetés comme n'étant pas torturés (car ils correspondent à des fautes de frappe, c'est-à-dire à des erreurs non intentionnelles) dans la phase d'évaluation.

Extraction et classification d'acronymes torturés

Étant donnés le corpus et ses caractéristiques associées, nous avons élaboré une chaîne de traitements permettant l'extraction et la classification d'acronymes, divisé en quatre tâches principales (représentées en Figure 3). Celui-ci sert de ligne de base pour cette tâche.

La tâche (I) consiste à extraire le contenu des articles au format PDF en TEI-XML par l'utilisation de *Grobid*[5], puis à en extraire le contenu textuel brut par l'utilisation de *BeautifulSoup4*[6]. Les tâches (II) et (III) sont étroitement liées car, pour chaque fichier texte précédemment généré, les fautes de frappe telles que les parenthèses dupliquées sont corrigées, puis les caractères spéciaux (c'est-à-dire les caractères de liaison et de contrôle) sont supprimés ; ce

5. `https://grobid.readthedocs.io/en/latest/`
6. `https://crummy.com/software/BeautifulSoup/`

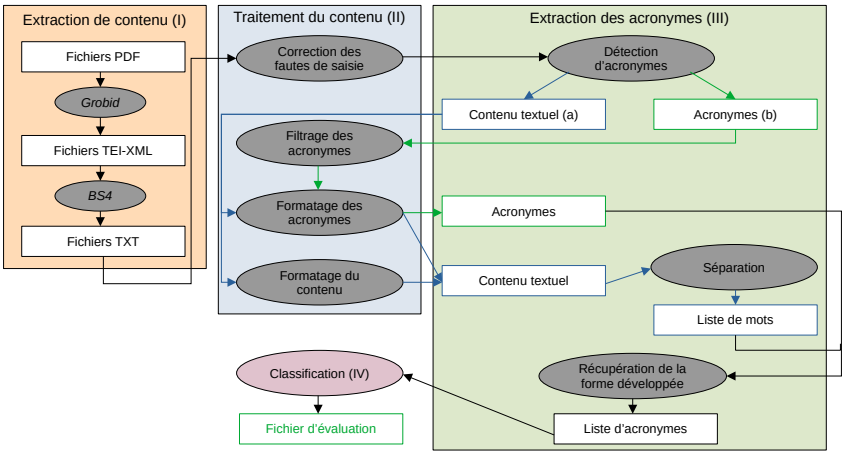

FIG. 4 – *Diagramme de traitement des acronymes. Le fonctionnement de la chaîne de traitements est divisé en quatre étapes, allant des l'extraction des données brutes depuis les fichiers PDF des articles à la création du fichier d'évaluation, en passant par l'extraction et la classification des acronymes. Les données textuelles brutes sont nettoyées en parallèle avec les acronymes extraits afin d'uniformiser les résultats obtenus.*

sont des tâches d'uniformisation de contenu. Une expression régulière est utilisée afin de détecter les acronymes, définis comme une suite de deux lettres capitales (ou plus), non séparées par un espace, le tout contenu entre parenthèses. Ces acronymes peuvent aussi contenir des points, des points d'interrogation, des tirets et esperluettes. À la fin de cette étape, le contenu textuel brut (III.a) et les acronymes (III.b) doivent être nettoyés. Ainsi, les acronymes sont filtrés afin de supprimer les faux-positifs tels que les références, et le texte brut est formaté selon qu'il s'agisse d'un acronyme ou non (par exemple les formes plurielles sont supprimées dans le cas d'un acronyme) puis séparé en liste de mots. La tâche (IV) consiste à classifier les acronymes, en suivant une séquence de règles de filtrage décrites dans l'algorithme 1, basées sur les caractéristiques communes aux acronymes torturés ou non.

Étant donné un acronyme extrait, une correspondance est effectuée sur les initiales (par exemple « *convolutional neural network (CNN)* », étape 1 de l'algorithme), en tenant également compte des mots de liaison (par exemple « *Moroccan agency of press (MAP)* », étape 2 de l'algorithme). Les mêmes comparaisons sont effectuées avec un décalage d'un mot (par exemple « *deep neural network models (DNN)* » et « *centers for disease control and prevention (CDC)* », étapes 3 et 4 de l'algorithme). Une correspondance est aussi effectuée sur les mots composés (par exemple « *chitosan (CS)* », « *electromyogram (EMG)* » et « *rectified linear unit (ReLU)* », étapes 5 et 6 de l'algorithme) et les formes réduites (par exemple « *residual network (ResNet)* », étape 6 de l'algorithme). Si aucune de ces correspondances n'est possible alors l'acronyme est considéré comme étant torturé. Enfin, un fichier d'évaluation est construit, contenant pour chaque acronyme, son étiquette et nombre d'occurrences tels qu'annotés dans le *silver* et tels qu'extraits par l'algorithme. Ce dernier est indépendant de toute liste d'acronymes de référence d'un quelconque domaine métier.

Données : un acronyme de longueur n précédé de $n + 1$ mots
Résultat : un triplet (forme développée, acronyme, étiquette)
$A \leftarrow$ l'acronyme;
$M \leftarrow$ l'acronyme découpé selon ses lettres capitales;
$P \leftarrow$ la liste des $n + 1$ mots avant l'acronyme;
$S \leftarrow$ la liste des $n + 1$ mots avant l'acronyme, sans les mots de liaison;
`// Étape 1`
si *les initiales de sousChaine(P, 1, longueur(P) - 1) = A* **alors**
  renvoyer (sousChaine(P, 1, longueur(P) - 1), A, non torturé);
fin
`// Étape 2`
si *les initiales de sousChaine(S, 1, longueur(S) - 1) = A* **alors**
  renvoyer (sousChaine(S, 1, longueur(S) - 1), A, non torturé);
fin
`// Étape 3`
si *les initiales de sousChaine(P, 2, longueur(P)) = A* **alors**
  renvoyer (sousChaine(P, 2, longueur(P)), A, non torturé);
fin
`// Étape 4`
si *les initiales de sousChaine(S, 2, longueur(S)) = A* **alors**
  renvoyer (sousChaine(S, 2, longueur(S)), A, non torturé);
fin
`// Étape 5`
pour tout *mot $p \in P$* **faire**
  si *les lettres de A se suivent dans p* **alors**
    renvoyer (p, A, non torturé);
  fin
fin
`// Étape 6`
si *les composantes de M sont alignées avec les mots de sousChaine(P, 2, longueur(P))* **alors**
  renvoyer (sousChaine(P, 2, longueur(P)), A) non torturé;
fin
renvoyer (P, A, torturé);

Algorithme 1 : Algorithme de classification d'un acronyme.

Benchmark

La tâche (III) obtient un rappel de 0,99, une précision de 0,89 et une F-mesure de 0,94. Nous avons obtenu pour la tâche (IV) un rappel de 0,95, une précision de 0,79, une F-mesure de 0,86 et un MCC de 0,78. Enfin, nous avons obtenu pour la tâche d'extraction d'acronymes torturés un rappel de 0,93, une précision de 0,62 et une F-mesure de 0,74. En complément, nous avons calculé les matrices de confusion de ces trois tâches (présentées en Figure 5) ainsi que la courbe ROC pour la tâche de classification des acronymes (décrite en Figure 6).

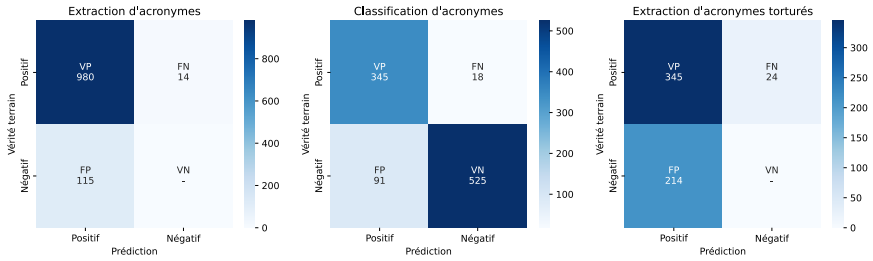

FIG. 5 – *Matrices de confusion pour les trois tâches (extraction, classification, extraction d'acronymes torturés). Nous constatons l'absence de VN dans la première et troisième tâche car ceux-ci n'ont pas été définis ; il y a aussi beaucoup de FP, peu importe la tâche.*

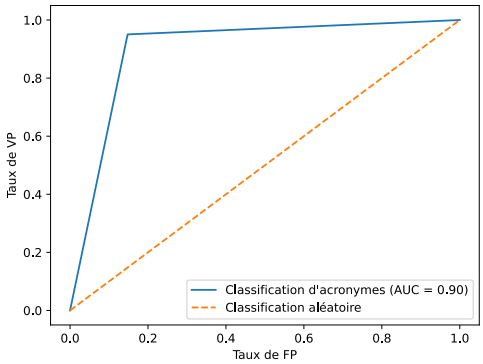

FIG. 6 – *Courbe ROC de la tâche de classification d'acronymes. Nous constatons une aire sous la courbe à 0,90, indiquant une bonne capacité discriminatoire.*

Étant donnés les résultats obtenus, l'évaluation de la tâche d'extraction d'acronymes semble satisfaisante, cependant ce n'est pas le cas des deux autres tâches. En effet, malgré un rappel supérieur à 0,9, les scores de précision indiquent un nombre trop important de FP, ces tâches ont donc tendance à surestimer les acronymes à traiter. Ceci pouvant être expliqué par les limites de notre méthode agnostique (basée sur les caractéristiques connues et communes aux acronymes torturés ou non) mais aussi par un nombre limité d'acronymes distincts dans notre corpus. Ce

dernier comprend un biais de représentativité au niveau des disciplines, avec une forte présence de termes d'informatique. Il devrait donc être enrichi, en effectuant des recherches en dehors des seuls articles dont le domaine métier est explicitement mentionné, mais aussi en y adjoignant d'autres corpus (par exemple celui proposé par (Lay et al., 2022)). Cela permettrait d'étudier davantage la sensibilité de notre algorithme d'extraction et de classification aux diverses subtilités induites par les formes variées d'acronymes. Nous avons constaté la présence d'un acronyme halluciné (« *bolster vector machine (BVM)* », une version torturée et non existante du « *support vector machine (SVM)* »), étiqueté comme étant légitime par notre algorithme. Il s'agit d'un cas anecdotique, cependant il serait intéressant d'étudier le contexte dans lequel a pu apparaître un tel acronyme. Aussi, le *silver* ayant été établi par une personne non polymathe, il se peut que des erreurs d'annotations soient présentes. Certains choix d'annotations peuvent être discutés, par exemple la forme développée du terme « *Bidirectional LSTM (BiLSTM)* » comporte à la fois un acronyme et un mot composé, qui sont deux caractéristiques distinctes.

Conclusion

Pour lutter contre la pollution de la littérature scientifique, nous avons défini une tâche d'extraction d'acronymes torturés : une forme spécifique de phrases torturées. Pour cela, nous avons constitué une collection de test, composée d'articles scientifiques en accès libre, dont nous avons extrait et annoté les acronymes, puis nous avons établi diverses métriques permettant d'évaluer cette tâche. Après avoir proposé une première méthode agnostique, nous avons obtenu une F-mesure de 0,74. Cependant, notre algorithme a tendance à trop surestimer les acronymes, amenant à un nombre trop important de faux-positifs, et par conséquent à une précision améliorable. De futurs travaux devraient être focalisés sur la production d'algorithmes plus performants, en incluant les décisions des experts dans la prise de décision algorithmique. Sur le long terme, il serait intéressant de mettre à disposition cette ligne de base au travers d'un défi TextMine [7] ou Kaggle [8]. Ces travaux pourraient aussi servir aux maisons d'édition, afin de filtrer leur chaîne éditoriale en amont, par l'identification et le signalement de potentiels problèmes aux éditeurs en charge d'organiser l'évaluation par les pairs. Nous avons constitué cette ligne de base afin que toute personne, désireuse de travailler sur l'extraction d'acronymes torturés, puisse s'y référer.

Disponibilité des données

Les données de test supportant cette étude sont disponibles sur Zenodo [9].

7. https://textmine.sciencesconf.org/
8. https://kaggle.com/
9. https://doi.org/10.5281/zenodo.10492230

Remerciements

Le projet NanoBubbles bénéficie d'un financement Synergy grant du Conseil de la recherche européen (European Research Council, ERC), dans le cadre du programme Horizon 2020 de l'Union Européenne, numéro de contrat 951393.

Références

Pennington, J., Socher, R. et Manning, C. (2014). *GloVe: Global Vectors for Word Representation*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'2014), ACL, Octobre 2014, p. 1532–1543 (DOI : https://doi.org/10.3115/v1/D14-1162).

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H et Mikolov, T. (2016). *FastText.zip: Compressing text classification models*. arXiv (DOI : https://doi.org/10.48550/arXiv.1612.03651)

Devlin, J., Chang, M.-W., Kenton, L. et Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv (DOI : https://doi.org/10.48550/arXiv.1810.04805).

Beltagy, I., LO, K. et Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. arXiv (DOI : https://doi.org/10.48550/arXiv.1903.10676).

Raffel, C., Shazerr, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W et Liu, P.-J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of Machine Learning Research, vol. 21 (140), 2020, p. 1–67 (URL : https://jmlr.org/papers/v21/20-074.html).

Biagioli, M. et Lippman, A. (2020). *Gaming the Metrics: Misconduct and Manipulation in Academic Research*. MIT Press, 2020, 307 p. (DOI : https://doi.org/10.7551/mitpress/11087.001.0001).

Barbour, B. et Stell, B.-M. (2020). *PubPeer: Scientific Assessment Without Metrics*. Gaming the Metrics : Misconduct and Manipulation in Academic Research, MIT Press, 2020, ch. 11 (DOI : https://doi.org/10.7551/mitpress/11087.003.0015).

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

Cabanac, G., Labbé, C. et Magazinov, A. (2022). *The 'Problematic Paper Screener' automatically selects suspect publications for post-publication (re)assessment*. CoRR, vol. abs/2107.06751 (DOI : https://doi.org/10.48550/arXiv.2210.04895).

Lay, P., Landschat, M. et Labbé, C. (2022). *Investigating the detection of Tortured Phrases in Scientific Literature*. In Proceedings of the Third Workshop on Scholarly Document Processing (SDP'2022), ACL, Octobre 2022, p. 32–36 (DOI : https://doi.org/10.48550/arXiv.2210.13024).

Kashnitsky, Y., Herrmannova, D., Waard, de, A., Tsatsaronis, G., Fennell, C. et Labbé, C. (2022). *Overview of the DAGPap22 Shared Task on Detecting Automatically Generated*

Scientific Papers. In Proceedings of the Third Workshop on Scholarly Document Processing (SDP'2022), ACL, Octobre 2022, p. 210–213 (URL : https://aclanthology.org/2022.sdp-1.4/).

Becker, J., Wahle, J.-P., Ruas, T. et Gipp, B. (2023). *Paraphrase Detection: Human vs. Machine Content*. arXiv (DOI : https://doi.org/10.48550/arXiv.2303.13989).

Summary

Public policies push researchers to steadily publish scientific articles in reputable journals. This pressure leads a minority of unscrupulous people to commit fraud, notably by resorting to tortured phrases to disguise plagiarism. Sometimes this type of content is not detected by the peer review process, leading to its publication. In order to decontaminate the scientific literature, we propose a tortured acronym extraction task to detect fraudulent publications. A benchmark is run using a corpus of 75 tortured scientific articles in open access, with an acronym extraction and classification pipeline. We obtained an F-score of 0.74 for the tortured acronym extraction task, which can be improved by enriching the dataset. We have created a baseline against which anyone wishing to improve this task can refer to.

Caractéristique	Classe		
	Non torturé	Torturé	Suspect
Fautes de frappe	-	-	Tru_-Positive (TP) – 47 acronymes de 29 sources distinctes
Même nombre d'initiales (1)	Pain Dysfunction Syndrome (PDS) – 398 acronymes de 68 sources distinctes	Concealed Markov Display (HMM) – 253 acronymes de 69 sources distinctes	-
Nombre différent d'initiales (2)	Centers for Diseases Control and Prevention (CDC) – 50 acronymes de 31 sources distinctes	Summed Up Direct Models (GLM) – 75 acronymes de 37 sources distinctes	-
Mots composés (3)	Intraocular Pressure (IOP) – 72 acronymes de 35 sources distinctes	Multidrug Safe (MDR) – 11 acronymes de 10 sources distinctes	-
Composés chimiques (4)	3-Aminopropyltriethoxysilane (3-APTES) – 43 acronymes de 5 sources distinctes	Diethylenetriamine Pentaacetic Ruinous (DTPA) – 4 acronymes de 2 sources distinctes	-
Formes réduites (5)	Residual Networks (ResNets) – 5 acronymes de 3 sources distinctes	Lingering Brain Organizations (ResNet) – 7 acronymes de 6 sources distinctes	-
Semi-acronymes	Gabor-HOG (GHOG) – 12 acronymes de 9 sources distinctes	Non-straight ARMA models (NARMA) – 12 acronymes de 9 sources distinctes	-
Nombres	Three-Dimensional (3D) – 2 acronymes de 2 sources distinctes	Fifth-age (5G) – 1 occurrence	-
Synonymes	-	Human PC Interface/Cooperation (HCI) - 1 occurrence	-
Suffixes supplémentaires	-	PC Assisted Determination (CADx) - 1 occurrence	-
Acronymes hallucinés	-	Bolster Vector Machine (BVM) - 1 occurrence	-

TAB. 1 – *Caractéristiques des acronymes par classe. Les comptages effectués ne tiennent pas compte du nombre d'occurrence d'un même acronyme dans une même source. Les sources peuvent comporter des acronymes de caractéristiques différentes. La classe des mots composés est distincte de celle du nombre différent d'initiales mais peut très bien englober celle des composés chimiques, une distinction a été faite car il pourrait être pertinent d'utiliser un lexique pour vérifier si un composé chimique est torturé ou non. Nous nous focalisons sur les caractéristiques (1) à (5), communes aux deux classes, qui représentent au total entre 95,63 et 97,59 % de celles-ci (voir Figure 3).*