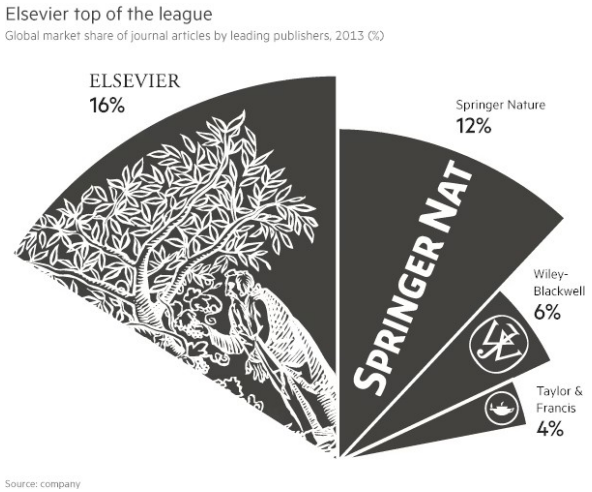
CNRS - Toulouse INP - UT3 - UT Capitole - UT2
Institut de Recherche en Informatique de Toulouse

Découverte d'abréviations torturées dans des publications scientifiques

Alexandre Clausse, Guillaume Cabanac, Pascal Cuxac, Cyril Labbé



https://www.ft.com/content/93138f3e-87d6-11e5-90de-f44762bf9896

voice recognition

deep neural network

Convolutional Neural Network. Artificial Neural Networks with numerous layers are termed as Deep Neural Networks or Deep Learning. It has been explored as one another key resource in recent years and has become quite well recognized in the literary community because of its efficiency to manage with huge amounts of data [17]. The most well-known profound neural network is the Convolutional Neural Networks (CNNs), which takes its name from operation of mathematical dimension from the matrixes termed convolution. Convolutional Neural Network (CNN) has various types of layers; it includes pooling, nonlinearity, and convolutional and fully connected layers. Convolutional Neural Network has pivotal outcomes over previous decades in an assortment of fields identified with design acknowledgment, from picture handling to voice acknowledgment [18]. The significant part of CNN is to get theoretical highlights when information proliferates towards the more profound layers. For instance, in picture characterization, the edge may be distinguished in the principal layers, and afterward the less difficult shapes in the subsequent layers, and afterward the more elevated level highlights.

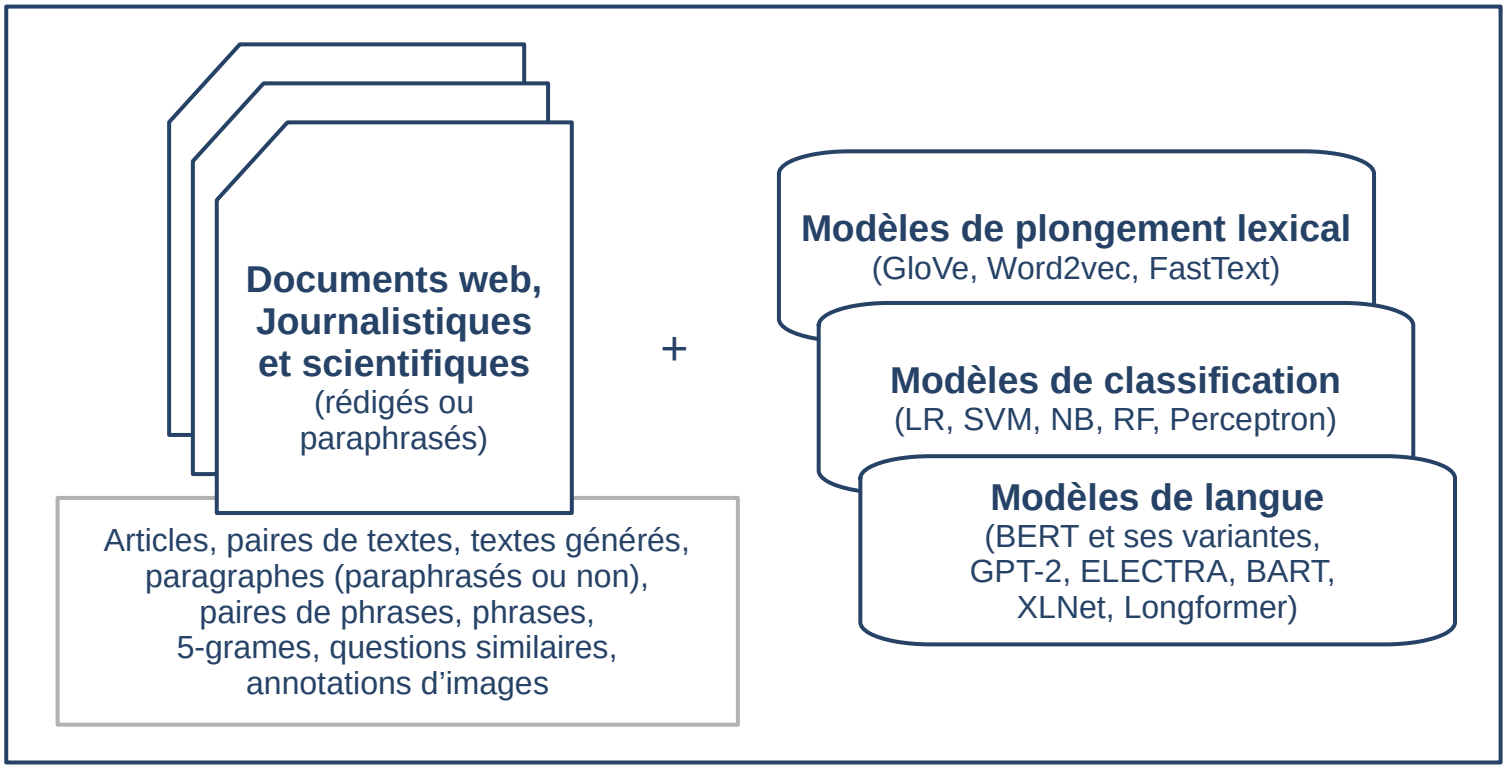
pattern recognition

deep layers

high-level features

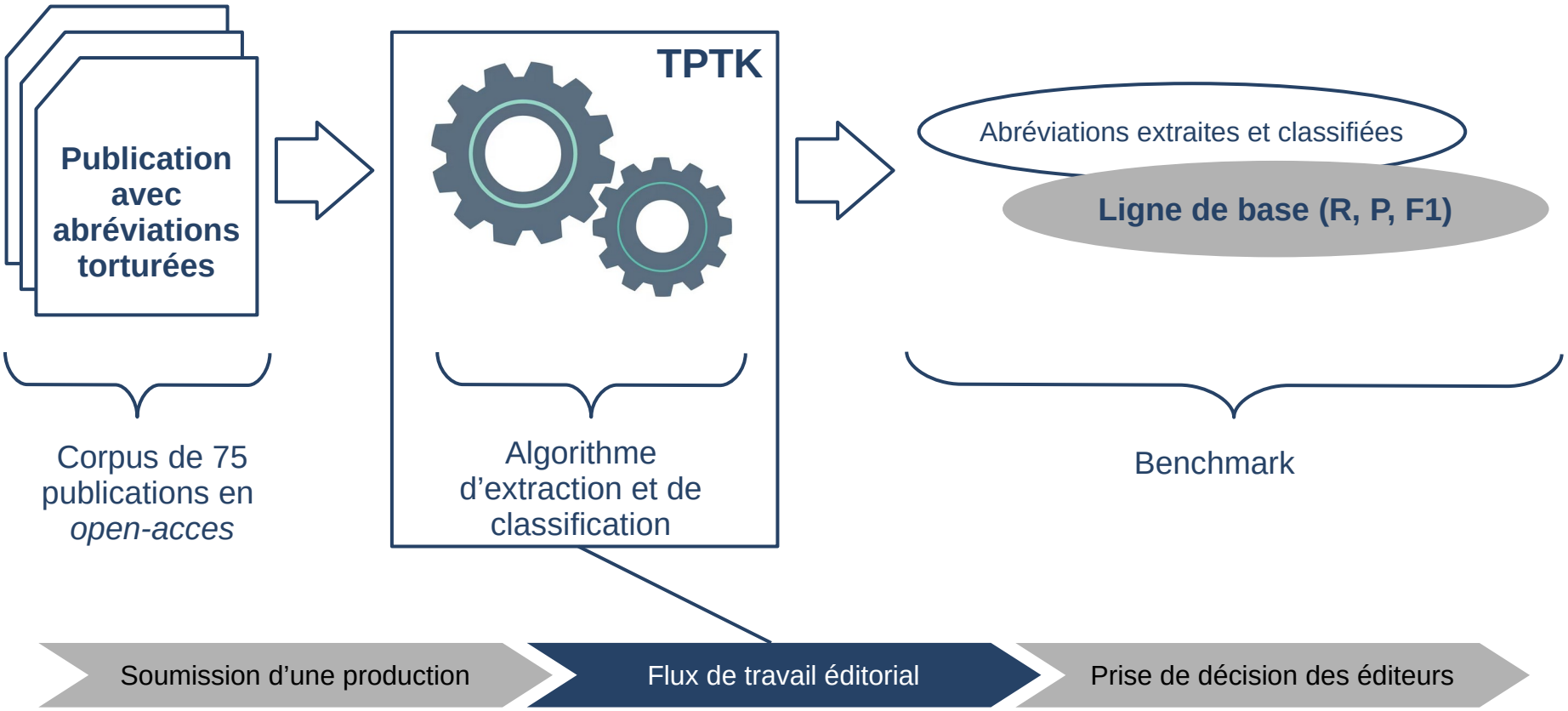
pubpeer.com/publications/99826B46D2BA62C9FA87CAF11AE0FD

Références : [3] et [4]



Références : [1], [2], [5], [7] à [11]

Abréviation torturée : *Bolster Vector Machine* (SVM) → *Support Vector Machine* (SVM)



Référence : [6]

Points forts de la contribution

- Approche indépendante du domaine métier
- Nécessite peu de ressources
- Traitement de documents PDF non publiés
- Pas besoin de liste d'expressions connues

Limites générales

- Propres aux études existantes
- Surestimation des abréviations
- Ne détecte pas les termes hallucinés et polysémiques
- Insensible aux néologismes

Pistes futures

- Données textuelles multilingues et de plusieurs domaines (modélisation thématique)
- Échantillons synthétiques paraphrasés et générés (à comparer avec du contenu authentique)
- Caractéristiques discriminatoires
- Identification des sources de plagiat (LMs + dictionnaires)
- Efficacité des LLMs (utilisation des RAG)

Abréviation hallucinée : *Bolster Vector Machine* (BVM) → *Support Vector Machine* (SVM)

Terme polysémique : *Profound Learning*

[1] Barbour B., Stell B. M., “PubPeer: Scientific Assessment Without Metrics”, in M. Biagioli, A. Lippman (ed.), Gaming the Metrics: Misconduct and Manipulation in Academic Research, The MIT Press, p. 149–155, 2020.
DOI : https://doi.org/10.7551/mitpress/11087.003.0015.

[2] Becker J., Wahle J. P., Ruas T., Gipp B., “Paraphrase Detection: Human vs. Machine Content”, Eighth Annual Conference on Advances in Cognitive Systems, Advances in Cognitive Systems, 2020.
URL : https://advancesincognitivesystems.github.io/acs/data/ACS2020_paper_28.pdf.

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

[4] Cabanac G., Labbe 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.

[5] Cabanac G., Labbe C., Magazinov A., “The ‘Problematic Paper Screener’ automatically selects suspect publications for post-publication (re)assessment”, 2022. Preprint arXiv.
DOI : https://doi.org/10.48550/arXiv.2210.04895.

[6] Clausse A., Cabanac G., Cuxac P., Labbe C., “Extraction d’acronymes tortures dans la litterature scientifique”, in P. Cuxac, C. Lopez (ed.), Atelier TextMine de la conference Extraction et Gestion des Connaissances (EGC) de 2024, Dijon, France, p. 27–37, 2024.
URL : https://hal.science/hal-04426448.

[7] Gehrmann S., Strobelt H., Rush A., “GLTR: Statistical Detection and Visualization of Generated Text”, in M. R. Costa-jussa, E. Alfonseca (ed.), 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.

[8] 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 (ed.), 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.

[9] Macbeth J. C., Chang E., Chen J. G., Grandic S., “A Broader Range for ‘Meaning the Same Thing’: Human Against Machine on Hard Paraphrase Detection Tasks”, 2020.
URL : https://api.semanticscholar.org/CorpusID:232247625.

[10] 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 (ed.), 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.

[11] Wahle J. P., Ruas T., Foltynek T., Meuschke N., Gipp B., “Identifying Machine-Paraphrased Plagiarism”, in M. Smits (ed.), 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.