« Vers la création d'une super-intelligence » : un corpus pour étudier les revendications des articles de TAL

Clémentine Bleuze[1] Fanny Ducel[2] Karën Fort[1] Maxime Amblard[1]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) Université Paris-Saclay, CNRS, LISN, F-91400 Orsay, France
clementine.bleuze@univ-lorraine.fr

MOTS-CLÉS : revendications, éthique, TAL pour le TAL.

KEYWORDS: claims, ethics, NLP4NLP.

1 Introduction

1.1 La question de l'exagération scientifique

L'exagération scientifique (*overclaiming* ou *spin* [1]) désigne le fait de présenter d'une manière excessivement avantageuse la portée d'un résultat issu de la recherche. Si ce phénomène ne s'accompagne pas nécessairement de comportements frauduleux, il porte néanmoins atteinte au principe d'intégrité scientifique et, par conséquent, à la qualité de la science et à la crédibilité de la communauté. Il n'est cependant pas aisé de définir objectivement ce qui constitue ou non une exagération scientifique : d'une part, les intentions des auteur·ices, leurs données de travail ainsi que leur démarche réelle restent inaccessibles par la simple lecture d'un article, d'autre part il peut être difficile de distinguer l'exagération de procédés rhétoriques inhérents à l'écriture de la recherche (Horton, 1995).

Dans le travail qui a mené à la constitution du corpus présenté ci-après, nous envisageons la notion d'exagération du point de vue des lecteur·ices qui pourraient se sentir trompé·es par certaines revendications [2]. Par ailleurs, si l'analyse de l'exagération porte habituellement sur les résultats principaux d'un article (voir par ex. Koroleva (2020)), nous considérons que ce phénomène peut impacter d'autres types de revendications, comme les annonces des contributions d'une étude ou de son impact sur la communauté [3].

1. Notion proche de l'*overclaiming* mais davantage utilisée dans le milieu médical, notamment au sujet des interprétations abusives de résultats de *RCTs* ; voir Koroleva (2020).

2. En particulier, nous ne présumons pas des intentions des auteur·ices au moment de l'écriture.

3. Par exemple, annoncer la création d'un modèle capable de traduire entre des paires de langues pour lesquelles aucune évaluation n'est pourtant présentée (exagération de contribution), ou laisser entendre qu'une étude à la portée assez limitée ouvre la voie à la création d'une « super-intelligence » (exagération d'impact).

1.2 Travaux en lien

L'étude de l'exagération scientifique s'inscrit dans le cadre plus large de l'étude des écrits scientifiques. Ce domaine bénéficie notamment de la disponibilité de bases de données ouvertes telles que l'*ACL Anthology* [4] (articles de linguistique computationnelle) ou *ArXiv* [5] (articles et pré-tirages dans de multiples domaines). Il peut aborder plusieurs axes, dont :

— **l'étude des écrits scientifiques comme genre textuel à part entière** : Les articles de recherche répondent en effet à des exigences relativement homogènes selon les disciplines, notamment en terme de structure et de portée rhétorique (Horton, 1995). La tâche de Zonage Argumentatif (*Argumentative Zoning*) (Teufel *et al.*, 1999) propose justement de classifier les phrases d'un article selon des catégories reflétant leur rôle argumentatif et rhétorique (par ex. CONTEXTE ou CONTRASTE). D'autres s'intéressent à un niveau plus fin, avec des catégories telles que RÉSULTATS ou IMPLICATIONS PRATIQUES dans les résumés d'articles (Stead *et al.*, 2019).

— **l'analyse réflexive de pratiques scientifiques d'un domaine donné** : Mariani *et al.* (2019) fournissent une étude longitudinale des données de l'*ACL Anthology* sur une période de 50 ans (1965-2015). Cela leur permet d'analyser une vaste gamme de phénomènes d'intérêt pour la communauté du TAL (dynamiques de citations entre auteur·ices, part des femmes dans les productions, évolution des sujets de recherche, etc.), une démarche qu'ils nomment « TAL pour le TAL » (*NLP4NLP*). Dans cette même perspective, Ducel (2022) étudie le degré de certitude des revendications (*claims*) d'articles de TAL, car cet indicateur pourrait constituer un *proxy* pour détecter l'exagération.

1.3 Contributions

Pour notre part, nous souhaitons faire communiquer ces deux axes. Premièrement, nous définissons une taxonomie des types de revendications contenues dans les articles de TAL [6]. Cette taxonomie est validée par des phases d'annotation itératives sur un large corpus d'articles en anglais que nous avons rassemblés. À l'aide d'annotations manuelles supplémentaires, nous prédisons automatiquement les catégories de revendications sur le reste de notre corpus avec un modèle *SciBERT* affiné (Beltagy *et al.*, 2019). Deuxièmement, en partant de l'hypothèse que la certitude peut constituer un *proxy* pour l'exagération, nous enrichissons le corpus avec des annotations en certitude à double niveau : phrase entière (*sentence-level*) et intra-phrase (*aspect-level certainty*), en utilisant les modèles de Pei & Jurgens (2021). Ceci nous permet d'esquisser de premières analyses sur la distribution de la certitude au sein des articles et des catégories de revendication. Notre contribution porte ici sur la publication de l'intégralité du

4. https://aclanthology.org/
5. https://arxiv.org/
6. C'est-à-dire, selon notre définition, des catégories de phrases pouvant être sujettes à l'exagération.

corpus annoté ainsi que du modèle utilisé pour l'annotation automatique.

2 Constitution et enrichissement d'un corpus de revendications en TAL

2.1 Collecte des données

L'*ACL Anthology* constitue une source incontournable d'articles de TAL publiés dans de multiples conférences (ACL, EMNLP, LREC, etc.), cependant nous décidons de l'enrichir avec des articles et pré-tirages publiés sur la plateforme *ArXiv* qui, bien que non relue par les pairs, héberge également une abondante production scientifique. Nous ré-utilisons le corpus ACL OCL (Rohatgi *et al.*, 2023) qui contient les méta-données et le contenu de 71 286 articles de l'*ACL Anthology* publiés entre 1952 et 2022, et extrayons le contenu des articles [7] *ArXiv* non redondants et en lien avec le TAL [8]. Nous obtenons les méta-données d'un ensemble de 105 101 articles, ainsi que 15 850 809 phrases issues de 87 767 d'entre eux [9].

phase	#categ.	anno.	#phrases annotées	#articles annotés	$\alpha(\uparrow)$	$\kappa(\uparrow)$ (min-max)
1	5	a1, a2, a3, a4	987 (a1-2) / 246 (a3-4)	10 (a1-2) / 4 (a3-4)	0,58	0,09-0,70
2	5	a1, a2, a5, a6	176	2	0,67	0,62-0,73
3	8	a1, a2	622	4	0,57	0,57
4	8	a1, a2	289	2	**0,81**	**0,81**

TABLE 1 – Statistiques des différentes phases d'annotation. En tout, 6 annotateur·ices (2 chercheur·euses, 3 doctorant·es et 1 étudiante de Master en TAL) ont pris part à la campagne.

2.2 Construction itérative d'une taxonomie de revendications *via* l'annotation manuelle d'articles de TAL

Partant d'une vision de l'exagération dépassant l'unique catégorie RÉSULTAT, nous construisons une taxonomie de catégories rhétoriques non-mutuellement exclusives que nous affinons progressivement (par ajout, suppression, ou fusion de catégories) lors de quatre phases d'annotations sur des articles issus de notre corpus (cf. Table 1). Les scores d'accord inter-annotateur·ices (α de Krippendorff, κ de Cohen) sont relevés à l'issue de chaque phase, et nourrissent les discussions sur les clarifications nécessaires dans le guide d'annotation. La

7. Nous utilisons désormais *article* au sens large, incluant les articles relus par les pairs et les pré-tirages.

8. Pour les détails du processus de collecte et pré-traitement des données, consulter le mémoire récapitulant l'ensemble de nos travaux sur le corpus ; voir Section 2.4.

9. Nous avons uniquement conservé le contenu des articles en anglais, et pour lesquels nous avons pu obtenir un fichier .xml bien formé.

catégorie	définitions et exemples (issus du corpus)
CONTEXTE	*Claims* donnant du contexte / des explications (*Most Semantic Role Labeling (SRL) approaches are supervised methods which require a significant amount of annotated corpus [...]*)
CONTRIBUTION	Déclaration de la nature des contributions de l'étude, objectifs, méthodes (*In this paper, we propose a Multi-Task Active Learning framework for Semantic Role Labeling [...]*)
PLAN (*outline*)	Phrases définissant la structure de l'article ou décrivant des figures (*Section 2 introduces the basic notions of ontologies [...]*)
RÉSULTAT	Résultats (expérimentaux ou non), analyses, discussions, opinions des auteur.ices (*According to our results, active learning is more efficient by using 12 % less of training data [...]*)
IMPACT	Impact anticipé ou observé sur les utilisateur.ices / la communauté (*[...] we expect that relational similarity measures will soon become widely used*)
DIRECTIONS	Intentions de poursuites du travail ou pistes de recherches futures (*As further work we propose to improve the taxonomy [...]*)
LIMITATION	Limitations anticipées ou observées, défauts et imperfections du travail présenté (*Inability to identify the named entity leads the system into this trap.*)
NON CLAIM	Toutes les autres phrases (notamment la méthodologie, ex : *We generated the VSM-WMTS results by adapting the VSM to the WMTS.*)

TABLE 2 – Définitions et exemples des catégories de notre taxonomie finale.

version finale de la taxonomie (phase 4) obtient les scores maximaux $\alpha = \kappa = 0,81$, ce qui confirme la pertinence des catégories retenues (cf. Table 2).

2.3 Annotation du corpus

Une fois la taxonomie validée par l'expérience, nous nous en servons pour annoter manuellement [10] 14 792 phrases issues de 158 articles de TAL : cela représente un total de 15 992 annotations (595 phrases ont plus d'une catégorie) et environ 27h de travail. La distribution des annotations collectées par catégorie est présentée dans la Table 3. Bien que certaines catégories restent peu représentées (IMPACT ne compte que 154 phrases), ce premier sous-corpus constitue un jeu de données de qualité, relativement représentatif de la littérature scientifique en TAL [11], suffisant pour entraîner un modèle de classification automatique (problème de classification multi-classes). En accord avec l'état de la recherche sur le sujet, nous testons des modèles de Régression Logistique et de Machines à Vecteurs de Support (*SVM*), ainsi que plusieurs modèles basés sur *BERT* (Devlin *et al.*, 2019) : nous obtenons les meilleurs résultats en termes de F-mesure (moyenne pondérée par classe : 0,89 sur les données de test) en affinant *SciBERT* [12] (Beltagy *et al.*, 2019). Nous utilisons ce modèle sur l'ensemble du corpus et obtenons des *silver labels* dont la distribution suit celle du sous-

10. Les annotations sont effectuées par les annotatrices a1 et a2 (cf. Table 1). Chaque phrase n'est annotée que par une annotatrice et les phrases d'un même article sont annotées séquentiellement.

11. Nous prenons garde d'équilibrer les sources (52.5 % *ACL Anthology*, 47.5 % *ArXiv*) et les périodes (15,2 % < 1994, 29,7 % entre 1994-2004, 27,2 % entre 2004-2014, 27,8 % > 2014) dans notre sélection. Notre but étant de ne pas négliger certaines sources ou périodes, nous notons toutefois la naïveté de cette distribution.

12. Nous utilisons le modèle `scibert_scivocab_uncased` disponible sur Huggingface.

	#anno.	CONT.	CONTR.	PLAN	RÉS.	IMP.	DIR.	LIM.	NC
ensemble	15 850 809	8,9 %	6,9 %	0,9 %	15,7 %	0,7 %	1,6 %	1,7 %	63,6 %
annoté	15 992	14,1 %	12,7 %	2,4 %	22,3 %	1,0 %	2,8 %	3,5 %	41,2 %

TABLE 3 – Part de chaque catégorie dans l'ensemble des annotations / dans les annotations manuelles.

corpus manuellement annoté (cf. Table 3). Une inspection naïve des prédictions du modèle nous confirme la pertinence globale de notre démarche. De même, nous utilisons les deux modèles d'évaluation de la certitude de Pei & Jurgens (2021) pour obtenir des *silver labels* de certitude au niveau de la phrase (*sentence-level*) et des aspects (*aspect-level*). L'ensemble de ces traitements sur le corpus représente plusieurs dizaines d'heures de calcul [13].

2.4 Disponibilité du corpus et perspectives

L'ensemble du corpus (méta-données, phrases annotées) et des fichiers associés (fichiers .xml, code, mémoire récapitulant l'ensemble de nos travaux sur le corpus [14]) sont disponibles librement sur *HuggingFace* [15] et *GitHub* [16], ainsi que le modèle *SciBERT* utilisé pour l'annotation automatique du corpus [17].

Il est important de préciser que ces données présentent des limitations. La fiabilité et l'interprétabilité des annotations ne sont pas assurées, notamment en ce qui concerne les annotations obtenues de manière automatique, et des erreurs de *parsing* des PDFs persistent. Notre proposition de taxonomie demeure également subjective, bien que nous ayons pris soin de la faire valider expérimentalement, et elle exclut par exemple les revendications liées à la méthodologie des articles. Nous croyons cependant que ce travail constitue un apport de valeur pour des chercheur·euses s'intéressant aux textes académiques (articles, pré-tirages), à la notion de certitude, ou pouvant tirer parti de nos annotations en catégories de revendications. Même pour des chercheur·euses d'autres disciplines, la réflexion autour de la difficile question de l'exagération peut amener à se questionner sur ses propres pratiques d'écriture et sur les impressions envoyées aux lecteur·ices et à la communauté. Nous n'excluons pas de poursuivre dans le futur l'amélioration et l'analyse du corpus, en lien avec la question de l'exagération qui a motivé notre étude initiale.

13. L'ensemble des calculs (entraînement des modèles et inférence) a été effectué sur la plateforme Grid5000.
14. Page répertoriant l'ensemble des liens cités ci-après : https://clementinebleuze.github.io/resources/
15. https://huggingface.co/datasets/ClementineBleuze/CNP
16. https://github.com/ClementineBleuze/claims-in-NLP
17. https://huggingface.co/ClementineBleuze/scibert_prefix_cont_ll_SEP

Remerciements

Nous souhaitons remercier le programme ORION de l'Université de Lorraine [18] ayant financé le stage de master duquel découlent les travaux présenté dans cet article (ANR-20-SFRI-0009). Ce travail a également bénéficié d'un accès aux ressources de la plateforme de calcul Grid5000.

Références

BELTAGY I., LO K. & COHAN A. (2019). Scibert : A pretrained language model for scientific text. (arXiv :1903.10676). arXiv :1903.10676 [cs].

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert : Pre-training of deep bidirectional transformers for language understanding. (arXiv :1810.04805). arXiv :1810.04805 [cs], DOI : 10.48550/arXiv.1810.04805.

DUCEL F. (2022). Analyse des claims dans les articles de traitement automatique des langues à l'aide d'une méthode par apprentissage non supervisé. Mémoire de master, Sorbonne Université. https://github.com/FannyDucel/FannyDucel.github.io/files/9536432/ducel-fanny-memoirem1.pdf.

HORTON R. (1995). The rhetoric of research. *BMJ*, **310**(6985), 985–987. DOI : 10.1136/bmj.310.6985.985.

KOROLEVA A. (2020). *Assisted authoring for avoiding inadequate claims in scientific reporting*. phdthesis, Université Paris-Saclay ; Universiteit van Amsterdam.

MARIANI J., FRANCOPOULO G. & PAROUBEK P. (2019). The nlp4nlp corpus (i) : 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, **3**. DOI : 10.3389/frma.2018.00036.

PEI J. & JURGENS D. (2021). Measuring sentence-level and aspect-level (un)certainty in science communications. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9959–10011, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.784.

ROHATGI S., QIN Y., AW B., UNNITHAN N. & KAN M.-Y. (2023). The ACL OCL corpus : Advancing open science in computational linguistics. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 10348–10361, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.640.

STEAD C., SMITH S., BUSCH P. & VATANASAKDAKUL S. (2019). Emerald 110k : A multidisciplinary dataset for abstract sentence classification. In M. MISTICA, M. PICCARDI

18. https://www.univ-lorraine.fr/lue/orion/

& A. MACKINLAY, Éds., *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, p. 120–125, Sydney, Australia : Australasian Language Technology Association.

TEUFEL S., CARLETTA J. & MOENS M. (1999). An annotation scheme for discourse-level argumentation in research articles. In H. S. THOMPSON & A. LASCARIDES, Éds., *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, p. 110–117, Bergen, Norway : Association for Computational Linguistics.