Simplification de Textes Scientifiques (et Rien de Plus) Rapport sur l'Action CLEF 2025 SimpleText

Liana Ermakova[1] Hosein Azarbonyad[2] Jan Bakker[3]
Benjamin Vendeville[1,4] Jaap Kamps[3]
(1) Université de Bretagne Occidentale, HCTI, Brest, France
(2) Elsevier, Amsterdam, The Netherlands
(3) University of Amsterdam, Amsterdam, The Netherlands
(4) Lab-STICC (UMR CNRS 6285), Brest, France
{liana.ermakova,benjamin.vendeville}@univ-brest.fr,
h.azarbonyad@elsevier.com, {j.bakker,kamps}@uva.nl

RÉSUMÉ

Ces dernières années, l'action SimpleText a rassemblé une communauté active de chercheurs en traitement du langage naturel (TLN) et en recherche d'information (RI) autour d'un objectif commun : améliorer l'accessibilité des textes scientifiques. Ses références en matière de recherche d'extraits scientifiques, de détection et d'explication de terminologies scientifiques, ainsi que de simplification de textes scientifiques sont désormais des standards. En 2025, nous introduisons cette année des changements majeurs dans l'organisation et les missions de l'action. L'action CLEF 2025 SimpleText proposera trois tâches principales. Tâche 1 sur *Simplification de texte* : *simplification de texte scientifique*. Tâche 2 sur *Créativité contrôlée* : *identifier et éviter les hallucinations*. Tâche 3 sur*SimpleText 2024 Revisité* : *tâches sélectionnées sur demande populaire*.

ABSTRACT

CLEF 2025 SimpleText Track : Simplify Scientific Text (and Nothing More)

This paper highlights the evolution and future directions of the Simple-Text Track at CLEF, which, over the last few years, fostered an active NLP and IR research community focused on improving access to scientific text. Its benchmarks on scientific passage retrieval, scientific terminology detection and explanation, and scientific text simplification have become standard references. After using a similar setup of the track in 2021-2024, we propose substantial modifications to the track's structure and tasks. The CLEF 2025 SimpleText track will contain the following three tasks. *Task 1 on Text Simplification : simplify scientific text. Task 2 on Controlled Creativity : identify and avoid hallucination. Task 3 on SimpleText 2024 Revisited : selected tasks by popular request.*

MOTS-CLÉS : Accès à l'information, recherche d'informations, simplification des textes, traitement du langage naturel, intelligence artificielle .

KEYWORDS: Information access, Information retrieval, Text simplification, Natural language processing, Artificial intelligence .

ARTICLE : **Accepté à** ECIR 2025 (Ermakova *et al.*, 2025).

1 Introduction

L'accès aux connaissances scientifiques reste limité pour le grand public en raison de leur complexité linguistique, notamment dans des domaines comme la recherche biomédicale, où l'enjeu est pourtant crucial. La simplification automatique des textes scientifiques, portée par l'essor de l'IA, vise à rendre ces informations plus accessibles. Depuis 2021, l'action SimpleText explore ces défis en développant des approches innovantes d'apprentissage automatique et d'évaluation des modèles de simplification. En 2025, après trois ans de recherche sur un même format, des évolutions majeures seront introduites pour améliorer l'accessibilité et la fiabilité de ces outils.

2 Tâches CLEF 2025 SimpleText

Tâche 1 : Simplification de texte La *Simplification de texte* vise la *simplification de texte scientifique*. Dans l'action CLEF 2025 SimpleText, nous introduisons Cochrane-auto, un corpus basé sur des résumés biomédicaux et leurs synthèses vulgarisées issues des revues Cochrane (Bakker & Kamps, 2024). Il étend la simplification scientifique avec des données parallèles authentiques et accessibles, intégrant fusion et réarrangement de phrases tout en respectant la structure discursive. Nous proposons deux tâches : la simplification au niveau des phrases (*Tâche 1.1*), et des documents (*Tâche 1.2*), avec une sous-tâche d'alignement de texte pour la Tâche 2. L'évaluation combine métriques automatiques (Davari *et al.*, 2024) et validation humaine.

Tâche 2 : Créativité contrôlée La tâche vise à *identifier et éviter les hallucinations*. En 2024, 47 % des soumissions SimpleText contenaient des erreurs dans au moins 10 % des phrases, et 19 % dépassaient 50 % (Ermakova *et al.*, 2024a). Nous analysons l'attribution des sources et détectons les hallucinations à partir d'un jeu de données aligné et étiqueté. Nous proposons deux tâches : *Tâche 2.1* : Identifier la génération créative et annotation des phrases selon leur fidélité à la source, avec et sans accès à celle-ci (Mickus *et al.*, 2024). Et *Tâche 2.2* : Produire un texte fidèle à sa source en limitant la génération créative, et comparaison de soumissions avec et sans attribution explicite des sources. L'évaluation combine métriques automatiques (Davari *et al.*, 2024) et validation humaine.

Tâche 3 : SimpleText 2024 Revisité L'objectif de la tâche *SimpleText 2024 Revisité* est d'effectuer les *tâches sélectionnées sur demande populaire*, et certaines activités des éditions CLEF 2024 SimpleText (Tâches 1, 2 et 4) pourraient être prolongées selon la demande. Les tâches sélectionnées seront hébergées sur CodaBench (`https://codabench.org/`). Les détails sont disponibles dans l'article LNCS de l'action *CLEF 2024 SimpleText* (Ermakova *et al.*, 2024b) et les publications CEUR des tâches *CLEF 2024 SimpleText : Tâche 1* (Sanjuan *et al.*, 2024), *Tâche 2* (Di Nunzio *et al.*, 2024), et *Tâche 4* (D'Souza *et al.*, 2024).

3 Conclusion

En s'appuyant sur l'expérience acquise les années précédentes, l'action SimpleText propose trois nouvelles tâches pour 2025.

Remerciements Benjamin Vendeville et Liana Ermakova sont financés par l'ANR (ANR-22-CE23-0019-01) et MaDICS CNRS (`https://www.madics.fr/ateliers/simpletext/`). Jan Bakker et Jaap Kamps sont soutenus par l'NWO (NWA #1518.22.105), NWO CI (#CISC.CC.016), l'Université d'Amsterdam (AI4FinTech) et l'ICAI (AI for Open Government Lab). Les opinions exprimées n'engagent pas les financeurs.

Références

BAKKER J. & KAMPS J. (2024). Cochrane-auto : An aligned dataset for the simplification of biomedical abstracts. In M. SHARDLOW, H. SAGGIO, F. ALVA-MANCHEGO, M. ZAMPIERI, K. NORTH, S. ŠTAJNER & R. STODDEN, Éds., *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (EMNLP-TSAR 2024)*.

DAVARI D., ERMAKOVA L. & KRESTEL R. (2024). Comparative analysis of evaluation measures for scientific text simplification. In A. ANTONACOPOULOS, A. HINZE, B. PIWOWARSKI, M. COUSTATY, G. M. DI NUNZIO, F. GELATI & N. VANDERSCHANTZ, Éds., *Linking Theory and Practice of Digital Libraries*, p. 76–91, Cham : Springer Nature Switzerland.

DI NUNZIO G. M., VEZZANI F., BONATO V., AZARBONYAD H., KAMPS J. & ERMAKOVA L. (2024). Overview of the CLEF 2024 SimpleText Task 2 : Identify and Explain Difficult Concepts. In (Faggioli *et al.*, 2024), p. 3129–3146.

D'SOUZA J., KABONGO S., GIGLOU H. B. & ZHANG Y. (2024). Overview of the CLEF 2024 SimpleText Task 4 : SOTA ? Tracking the State-of-the-Art in Scholarly Publications. In (Faggioli *et al.*, 2024), p. 3163–3173.

ERMAKOVA L., AZARBONYAD H., BAKKER J., VENDEVILLE B. & KAMPS J. (2025). CLEF 2025 SimpleText Track : Simplify Scientific Text (and Nothing More). In C. HAUFF & OTHERS, Éds., *ECIR 2025 : Proceedings of the 47th European Conference on Information Retrieval*, volume 15576 de *Lecture Notes in Computer Science* : Springer. DOI : 10.1007/978-3-031-88720-8_63.

ERMAKOVA L., LAIMÉ V., MCCOMBIE H. & KAMPS J. (2024a). Overview of the CLEF 2024 SimpleText Task 3 : Simplify Scientific Text. In (Faggioli *et al.*, 2024), p. 3147–3162.

ERMAKOVA L., SANJUAN E., HUET S., AZARBONYAD H., DI NUNZIO G. M., VEZZANI F., D'SOUZA J. & KAMPS J. (2024b). Overview of the CLEF 2024 SimpleText track : Improving access to scientific texts for everyone. In L. GOEURIOT, G. Q. PHILIPPE MULHEM, D. SCHWAB, L. SOULIER, G. M. D. NUNZIO, P. GALUŠČÁKOVÁ, A. G. S. DE HERRERA, G. FAGGIOLI & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science : Springer.

FAGGIOLI G., FERRO N., GALUŠČÁKOVÁ P. & GARCÍA SECO DE HERRERA A., Éds. (2024). *Working Notes of CLEF 2024 : Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings. CEUR-WS.org.

MICKUS T., ZOSA E., VAZQUEZ R., VAHTOLA T., TIEDEMANN J., SEGONNE V., RAGANATO A. & APIDIANAKI M. (2024). SemEval-2024 task 6 : SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In A. K. OJHA, A. S. DOĞRUÖZ, H. TAYYAR MADABUSHI, G. DA SAN MARTINO, S. ROSENTHAL & A. ROSÁ, Éds., *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, p. 1979–1993, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.semeval-1.273.

SANJUAN E., HUET S., KAMPS J. & ERMAKOVA L. (2024). Overview of the CLEF 2024 SimpleText Task 1 : Retrieve Passages to Include in a Simplified Summary. In (Faggioli *et al.*, 2024), p. 3115–3128.