Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Quand des Non-Experts Recherchent des Textes Scientifiques Rapport sur l'action CLEF 2023 SimpleText

Liana Ermakova[1] Stéphane Huet[2] Éric SanJuan[2]
Olivier Augereau[3] Hosein Azarbonyad[4] Jaap Kamps[5]

(1) Université de Bretagne Occidentale, HCTI, 29200 Brest, France
(2) Avignon Université, LIA, France
(3) ENIB, Lab-STICC UMR CNRS 6285, France
(4) Elsevier, The Netherlands
(5) University of Amsterdam, The Netherlands
liana.ermakova@univ-brest.fr, contact@simpletext-project.com

RÉSUMÉ

Le grand public a tendance à éviter les sources fiables telles que la littérature scientifique en raison de leur langage complexe et du manque de connaissances nécessaires. Au lieu de cela, il s'appuie sur des sources superficielles, trouvées sur internet ou dans les médias sociaux et qui sont pourtant souvent publiées pour des raisons commerciales ou politiques, plutôt que pour leur valeur informative. La simplification des textes peut-elle contribuer à supprimer certains de ces obstacles à l'accès ? Cet article présente l'action « CLEF 2023 SimpleText » qui aborde les défis techniques et d'évaluation de l'accès à l'information scientifique pour le grand public. Nous fournissons des données réutilisables et des critères de référence pour la simplification des textes scientifiques et encourageons les recherches visant à faciliter à la compréhension des textes complexes.

ABSTRACT

What Happens if Non-Experts Search Scientific Texts ? CLEF 2023 SimpleText track report

Users tend to stay away from credible sources such as scientific literature due to their intricate language or their lack of prior knowledge. Rather, they depend on shallow sources from the web or social media that can be published for economic or political motives rather than for their enlightening value. Text simplification might remove some of these barriers. This paper presents the report on the CLEF 2023 SimpleText track aiming to encourage the research on automatic simplification of scientific texts by providing reusable data and benchmarks.

MOTS-CLÉS : Textes scientifiques, simplification, recherche d'information.

KEYWORDS: Scientific texts, simplification, information retrieval.

1 Introduction

Les textes scientifiques tels que les articles de recherche sont difficiles à comprendre pour le grand public et même pour les scientifiques qui ne sont pas spécialisés dans le domaine en question. L'action CLEF 2023 SimpleText (Ermakova *et al.*, 2023) se concentre sur la simplification des textes scientifiques en combinant des aspects de la recherche d'information et du traitement automatique du

Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023, pages 27–33.

langage naturel (TALN).

La complexité des textes, les niveaux de lecture et la simplification des textes en général sont étudiés depuis longtemps dans les domaines de la linguistique, des sciences de l'éducation et du TALN. L'amélioration de la compréhensibilité des textes reste un défi, car il est difficile de définir le résultat souhaitable de la simplification (Grabar & Saggion, 2022). Les scores de lisibilité traditionnels se limitent à la longueur des mots ou des phrases, tandis que les mesures basées sur le chevauchement du vocabulaire ne tiennent pas compte de la distorsion de l'information.

Récemment, la simplification des textes a suscité un intérêt croissant. L'atelier sur le traitement des documents scientifiques [1] s'est adressé à un public spécialisé dans le TALN (Chandrasekaran *et al.*, 2020), en accueillant des tâches sur le résumé de documents scientifiques. Lors de l'EMNLP 2022, le projet TSAR (*Text Simplification, Accessibility, and Readability*) [2] a accueilli une tâche de simplification lexicale, et TermEval 2020 a exécuté une tâche partagée sur l'extraction automatique de termes (Rigouts Terryn *et al.*, 2020). L'action SimpleText que nous proposons ne se limite pas à la simplification lexicale et grammaticale.

SimpleText vise à améliorer l'accès à la connaissance scientifique pour le grand public, en fournissant des données réutilisables et des critères de référence pour la simplification des textes pour réduire les obstacles à la compréhension des textes complexes. Contrairement aux travaux précédents, nous nous concentrons sur (1) la sélection d'informations adaptées au grand public, (2) la recherche de concepts difficiles, notamment de mots, d'abréviations, etc. qui doivent être expliqués, et (3) l'évaluation de la distorsion de l'information susceptible de se produire au cours du processus de simplification.

SimpleText est basée sur la série suivante d'actions (Ermakova *et al.*, 2022c) : (1) sélectionner les informations à inclure dans un résumé simplifié ; (2) décider si les informations sélectionnées sont suffisantes et compréhensibles ou fournir des connaissances de base si ce n'est pas le cas ; et (3) améliorer la lisibilité du texte. Il en résulte trois tâches :

- ***Tâche 1: What is in, or out?*** *Selectionner des passages à inclure dans un résumé simplifié,*
- ***Tâche 2: What is unclear?*** *Identifier et expliquer les concepts complexes,*
- ***Tâche 3: Rewrite this!*** *Simplifier, réécrire un texte scientifique.*

Une quatrième tâche ouverte accepte également toute soumission qui utilise nos données d'une autre manière. Dans la suite de cet article, nous allons présenter chacune des trois premières tâches.

Pour cette nouvelle édition, nous enrichirons les données produites par les éditions précédentes (Ermakova *et al.*, 2022c) en y ajoutant des labels. Nous mettrons également l'accent sur l'évaluation automatique sur des données de test réutilisables.

2 Tâche 1 : Selectionner des passages à inclure dans un résumé simplifié

Cette tâche vise à extraire des passages qui peuvent aider à comprendre un article de vulgarisation scientifique à partir d'un large corpus de résumés académiques et de métadonnées bibliographiques. Les passages pertinents doivent se rapporter à l'un des sujets de l'article source.

Données. Nous utilisons les articles de vulgarisation scientifique comme sources pour les types de

1. https://sdproc.org/2022/sharedtasks.html
2. https://taln.upf.edu/pages/tsar2022-st/

sujets qui intéressent le grand public et comme validation du niveau de lecture qui leur convient. Le corpus principal est un vaste ensemble de résumés scientifiques et de métadonnées associées, couvrant le domaine de l'informatique et de l'ingénierie. Nous réutilisons la collection de résumés universitaires du Citation Network Dataset (12e version publiée en 2020) [3] (Tang *et al.*, 2008). Cette collection a été extraite de DBLP, ACM, MAG (Microsoft Academic Graph) et d'autres sources. Il contient : 4 894 083 références bibliographiques publiées avant 2020, 4 232 520 résumés en anglais, 3 058 315 auteurs avec leurs affiliations et 45 565 790 citations ACM. Nous fournissons un index ElasticSearch pour permettre aux participants de retrouver des passages ou des résumés à l'aide de BM25 (Robertson *et al.*, 2009). Grâce à une API, des requêtes peuvent être effectuées sur le contenu textuel des résumés ainsi que sur la paternité des auteurs. Ainsi, l'ensemble de données partagées fournit : a) le contenu des résumés des documents ; b) les auteurs des documents pour l'analyse des coauteurs ; c) la relation de citation entre les documents pour l'analyse des co-citations ; et d) les citations par auteur pour l'analyse du facteur d'impact des auteurs.

Les articles de presse utilisés s'adressent à un public général et proviennent de deux sources : *The Guardian*, un journal d'audience internationale destiné au grand public et comportant une section technologique, et *Tech Xplore*, [4] un site web qui participe au réseau Science X en se focalisant sur les progrès de l'ingénierie et de la technologie. Chacun de ces articles de vulgarisation scientifique représente un sujet général qui doit être analysé pour extraire les informations scientifiques pertinentes du corpus. Nous fournissons les URL des articles originaux, le titre et le contenu textuel de chaque article de vulgarisation scientifique en tant que sujet général. Chaque thème général a été également enrichi d'un ou plusieurs mots-clés spécifiques extraits manuellement de leur contenu, ce qui crée une tâche classique de recherche d'informations consistant à classer les passages ou les résumés en réponse à une requête. Dans l'édition précédente, 40 articles, 20 de chaque source, ont été mis à disposition (Ermakova *et al.*, 2022c). Nous prévoyons de l'étendre à 10 autres sujets utilisés comme ensemble de test. Les qrels 2022 couvrent de nombreux sujets (31) et requêtes (67), mais avec une profondeur limitée en documents annotés. En 2023, nous augmenterons la profondeur avec au moins 50 résumés jugés par requête.

Évaluation. La pertinence thématique n'a été évaluée l'année dernière qu'avec une note de 0 à 5 sur le degré de pertinence par rapport à l'article original. Si cette grande échelle permet de mesurer la proximité du résumé extrait avec le mot-clé, le titre ou le contenu textuel, d'autres facettes, pourtant importantes dans le contexte de la simplification des textes, manquaient à l'appel. En 2023, nous continuerons à évaluer la pertinence thématique, mais aussi la complexité du texte (à l'aide de mesures de lisibilité et d'une comparaison avec des scores attribués manuellement) et l'autorité de la source (à l'aide de mesures de l'impact académique).

La collection de test fournie sera simplifiée en trois notes sur une échelle de 0 à 2 :

- **Pertinence du sujet** : Non pertinent (0), pertinent (1), très pertinent (2) ;
- **Complexité du texte** : Facile (0), difficile (1), très difficile (2) ;
- **Crédibilité de la source** : Faible (0), moyenne (1), forte crédibilité (2).

Tout en continuant à utiliser un classement sur la pertinence à l'aide du NDCG, les deux autres critères permettront de comparer les systèmes sur d'autres aspects à prendre en compte.

3. https://www.aminer.cn/citation
4. https://techxplore.com/

3 Tâche 2 : Identifier et expliquer les concepts complexes

L'objectif de cette tâche est de déterminer quels concepts dans les résumés scientifiques nécessitent une explication et une mise en contexte afin d'aider le lecteur à comprendre le texte scientifique. L'identification des mots complexes et la simplification lexicale sont les approches les plus populaires pour évaluer et réduire la complexité (Cruz *et al.*, 2019; Yimam *et al.*, 2018; Kochmar *et al.*, 2020). Dans le cadre d'une requête, certains concepts clés doivent être contextualisés avec une définition, un exemple ou des cas d'utilisation plus faciles à comprendre pour le lecteur. Des recherches sont en cours à ce sujet, en générant des définitions d'une complexité contrôlable (August *et al.*, 2022).

Nous demandons aux participants d'identifier ces concepts et de fournir des explications utiles et compréhensibles. La tâche comporte deux étapes : (1) retrouver jusqu'à cinq termes difficiles dans un passage donné d'un résumé scientifique ; (2) fournir une explication de ces termes difficiles (par exemple, définition, déchiffrage d'abréviation, etc.).

Données. Le corpus de la Tâche 2 est basé sur les phrases des résumés les mieux classés pour les requêtes de la Tâche 1. Pour la première étape de la tâche, c'est-à-dire l'extraction des termes difficiles, nous utiliserons les données d'entraînement recueillies en 2022 (Ermakova *et al.*, 2022c). En ce qui concerne les données de test, nous fournirons des passages supplémentaires provenant des résumés DBLP comme dans la Tâche 1.

Pour la deuxième étape de la tâche, nous fournirons des données d'entraînement supplémentaires pour la génération de définitions, extraites d'un corpus beaucoup plus large d'articles en texte intégral. Ces données de formation contiennent des paires de $< phrase, concept >$ et un label par paire est fournie. Une paire indique si la phrase fournit une bonne définition du concept ou non. Les échantillons de cet ensemble de données sont extraits de livres et d'articles publiés dans ScienceDirect [5]. En plus de cet ensemble de données, les participants sont encouragés à utiliser des ensembles de données existants extraits d'autres ressources, tels que l'ensemble de données de la CMT (Navigli & Velardi, 2010) pour entraîner le modèle de génération de définitions.

Évaluation Comme en 2022, nous évaluerons la détection de concepts complexes en fonction de leur complexité et de la portée des concepts détectés (Ermakova *et al.*, 2022c). Pour l'explication des termes difficiles, l'ensemble d'évaluation contiendra 1 000 concepts et leurs définitions extraites par des experts en la matière. Nous évaluerons automatiquement les explications fournies en les comparant à des références (par exemple ROUGE, similarité cosinus, etc.). Nous évaluerons manuellement les explications fournies au niveau de l'utilité par rapport à une requête et de la complexité pour un public général. Les explications fournies peuvent prendre différentes formes : définition, décryptage d'abréviations, exemples, cas d'utilisation, etc.

4 Tâche 3 : Simplifier, réécrire un texte scientifique

L'objectif de cette tâche est de fournir une version simplifiée des phrases extraites des résumés scientifiques. Les participants recevront les articles et requêtes de vulgarisation scientifique ainsi que les résumés correspondants d'articles scientifiques, divisés en phrases individuelles.

Données. La Tâche 3 utilise un corpus constitué de phrases issues des résumés les mieux classés

5. https://www.sciencedirect.com/

pour les requêtes de la Tâche 1, complétées par des données d'entraînement supplémentaires dans le domaine de la santé. Nos données d'entraînement sont un corpus véritablement parallèle de phrases directement simplifiées (648 phrases pour l'instant) provenant de résumés scientifiques du DBLP Citation Network Dataset pour le sujet *Computer Science* et d'articles de Google Scholar et PubMed sur *la santé et la médecine* (Ermakova *et al.*, 2021, 2022a,c,b). Ces passages de texte ont été simplifiés soit par des étudiants en master de rédaction technique et de traduction, soit par un expert du domaine (un informaticien) et un traducteur professionnel (de langue maternelle anglaise) travaillant ensemble (Ermakova *et al.*, 2022a,c,b).

Évaluation. En 2023, nous mettrons l'accent sur les mesures d'évaluation automatique à grande échelle (SARI, ROUGE, compression, lisibilité) qui fournissent une collection réutilisable de tests. Elles seront complétées par une évaluation humaine d'autres aspects, essentielle pour une analyse plus approfondie. Comme en 2022, nous évaluerons la qualité des simplifications au niveau du vocabulaire et de la syntaxe ainsi que les erreurs (syntaxe incorrecte ; anaphore non résolue due à la simplification ; répétition/itération inutile ; erreurs d'orthographe, de typographie ou de ponctuation) (Ermakova *et al.*, 2022c). Plutôt que de nous concentrer uniquement sur cette évaluation, qui est similaire à celle utilisée dans des travaux antérieurs (Štajner *et al.*, 2022), nous examinerons également les résultats du point de vue de la distorsion de l'information qui peut survenir au cours du processus de simplification, avec un niveau de gravité comme suit : Style (1) ; Insertion de détails inutiles par rapport à une requête (1) ; Redondance (sans chevauchement lexical) (2) ; Insertion d'informations fausses ou non étayées (3) ; Omission de détails essentiels par rapport à une requête (4) ; Généralisation excessive (5) ; Simplification excessive (5) ; Changement de sujet (5) ; Contre-sens / contradiction (6) ; Ambiguïté (6) ; Absurdité (7).

5 Conclusions

Cet article décrit la mise en place de l'action CLEF 2023 SimpleText, qui contient trois tâches interconnectées sur la simplification des textes scientifiques. Dans le cadre du projet SimpleText, nous avons déjà publié des corpus de taille conséquente et des données étiquetées manuellement :

- un corpus de plus de 4 millions de résumés scientifiques utilisable pour la vulgarisation scientifique,
- des termes scientifiques utilisés dans des résumés scientifiques, avec des scores de difficulté attribués manuellement,
- un corpus parallèle de phrases simplifiées manuellement à partir de la littérature scientifique,
- un corpus parallèle de phrases avec différents types de distorsion de l'information et de niveau de simplification.

Pour plus de détails, il est possible de consulter le site (`http://simpletext-project.com`).

Remerciements

Ce travail n'aurait pas été possible sans le soutien de nombreuses personnes et le groupe de recherche MaDICS. Cette recherche a été financée en tout ou partie, par l'Agence Nationale de la Recherche (ANR) au titre du projet « ANR-22-CE23-0019-01 ».

Références

AUGUST T., REINECKE K. & SMITH N. A. (2022). Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 8298–8317.

CHANDRASEKARAN M. K., FEIGENBLAT G., FREITAG D., GHOSAL T., HOVY E., MAYR P., SHMUELI-SCHEUER M. & DE WAARD A. (2020). Overview of the first workshop on scholarly document processing (SDP). In *Proceedings of the First Workshop on Scholarly Document Processing*, p. 1–6, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.sdp-1.1.

CRUZ F., COUSTATY M., AUGEREAU O., KISE K. & JOURNET N. (2019). An interactive recommendation system for 2nd language vocabulary learning-vocabulometer 2.0. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 3, p. 28–32 : IEEE.

ERMAKOVA L., BELLOT P., BRASLAVSKI P., KAMPS J., MOTHE J., NURBAKOVA D., OVCHINNIKOVA I. & SANJUAN E. (2021). Overview of SimpleText 2021 - CLEF Workshop on Text Simplification for Scientific Information Access. In K. S. CANDAN, B. IONESCU, L. GOEURIOT, B. LARSEN, H. MÜLLER, A. JOLY, M. MAISTRO, F. PIROI, G. FAGGIOLI & N. FERRO, Éds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, p. 432–449, Cham : Springer International Publishing.

ERMAKOVA L., BELLOT P., KAMPS J., NURBAKOVA D., OVCHINNIKOVA I., SANJUAN E., MATHURIN E., ARAÚJO S., HANNACHI R., HUET S. & POINSU N. (2022a). Automatic Simplification of Scientific Texts : SimpleText Lab at CLEF-2022. In M. HAGEN, S. VERBERNE, C. MACDONALD, C. SEIFERT, K. BALOG, K. NØRVÅG & V. SETTY, Éds., *Advances in Information Retrieval*, volume 13186, p. 364–373. Cham : Springer International Publishing.

ERMAKOVA L., OVCHINNIKOVA I., KAMPS J., NURBAKOVA D., ARAÚJO S. & HANNACHI R. (2022b). Overview of the CLEF 2022 SimpleText Task 3 : Query biased simplification of scientific texts. CEUR Workshop Proceedings.

ERMAKOVA L., SANJUAN E., HUET S., AUGEREAU O., AZARBONYAD H. & KAMPS J. (2023). CLEF 2023 SimpleText Track. In J. KAMPS, L. GOEURIOT, F. CRESTANI, M. MAISTRO, H. JOHO, B. DAVIS, C. GURRIN, U. KRUSCHWITZ & A. CAPUTO, Éds., *Advances in Information Retrieval*, p. 536–545, Cham : Springer Nature Switzerland.

ERMAKOVA L., SANJUAN E., KAMPS J., HUET S., OVCHINNIKOVA I., NURBAKOVA D., ARAÚJO S., HANNACHI R., MATHURIN É. & BELLOT P. (2022c). Overview of the CLEF 2022 SimpleText Lab : Automatic simplification of scientific texts. In A. BARRÓN-CEDEÑO, G. D. S. MARTINO, M. D. ESPOSTI, F. SEBASTIANI, C. MACDONALD, G. PASI, A. HANBURY, M. POTTHAST, G. FAGGIOLI & N. FERRO, Éds., *CLEF'22 : Proceedings of the Thirteenth International Conference of the CLEF Association*, Lecture Notes in Computer Science : Springer.

GRABAR N. & SAGGION H. (2022). Evaluation of automatic text simplification : Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 453–463.

KOCHMAR E., GOODING S. & SHARDLOW M. (2020). Detecting multiword expression type helps lexical complexity assessment. In *LREC 2020 : Proceedings of the 12th Conference on Language Resources and Evaluation*.

NAVIGLI R. & VELARDI P. (2010). Learning word-class lattices for definition and hypernym extraction. In *ACL*, p. 1318–1327.

RIGOUTS TERRYN A., HOSTE V., DROUIN P. & LEFEVER E. (2020). Termeval 2020 : Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, p. 85–94 : European Language Resources Association (ELRA).

ROBERTSON S., ZARAGOZA H. *et al.* (2009). The probabilistic relevance framework : Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, **3**(4), 333–389.

TANG J., ZHANG J., YAO L., LI J., ZHANG L. & SU Z. (2008). ArnetMiner : Extraction and mining of academic social networks. In *KDD'08*, p. 990–998.

YIMAM S. M., BIEMANN C., MALMASI S., PAETZOLD G., SPECIA L., ŠTAJNER S., TACK A. & ZAMPIERI M. (2018). A report on the complex word identification shared task 2018. In *The 13th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL2018 Workshops)*.

ŠTAJNER S., SHEANG K. C. & SAGGION H. (2022). Sentence Simplification Capabilities of Transfer-Based Models.