CLEF 2024 SimpleText Track

Improving Access to Scientific Texts for Everyone

Ermakova, L.; SanJuan, E.; Huet, S.; Azarbonyad, H.; Di Nunzio, G.M.; Vezzani, F.; D'Souza, J.; Kabongo, S.; Giglou, H.B.; Zhang, Y.; Auer, S.; Kamps, J.

DOI
10.1007/978-3-031-56072-9_4

Publication date
2024

Document Version
Final published version

Published in
Advances in Information Retrieval

CLEF 2024 SimpleText Track

Improving Access to Scientific Texts for Everyone

Liana Ermakova[1(✉)], Eric SanJuan[2], Stéphane Huet[2], Hosein Azarbonyad[3], Giorgio Maria Di Nunzio[4], Federica Vezzani[4], Jennifer D'Souza[5], Salomon Kabongo[6], Hamed Babaei Giglou[5], Yue Zhang[7], Sören Auer[5], and Jaap Kamps[8]

[1] Université de Bretagne Occidentale, HCTI, Brest, France
liana.ermakova@univ-brest.fr
[2] Avignon Université, LIA, Avignon, France
[3] Elsevier, Amsterdam, The Netherlands
[4] University of Padua, Padua, Italy
[5] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
[6] L3S Research Center, Leibniz University of Hannover, Hanover, Germany
[7] Technische Universität Berlin, Berlin, Germany
[8] University of Amsterdam, Amsterdam, The Netherlands
https://simpletext-project.com

Abstract. Everyone agrees on the importance of objective scientific information. However, relevant scientific documents tend to be inherently difficult to find and understand either because of intricate terminology or the potential absence of prior knowledge among their readers. Can we improve accessibility for everyone? This paper introduces the SimpleText Track at CLEF 2024, addressing the technical and evaluation challenges associated with making scientific information accessible to a wide audience, including students and non-experts. We provide appropriate reusable data and benchmarks for scientific text summarization and simplification. The CLEF 2024 SimpleText track is based on four interrelated tasks: Task 1 *Content Selection*: retrieving passages to include in a simplified summary. Task 2 *Complexity Spotting*: identifying and explaining difficult concepts. Task 3 *Text Simplification*: simplify scientific text. Task 4 *SOTA?*: tracking the state-of-the-art in scholarly publications.

Keywords: Scientific text simplification · Information extraction · Information retrieval · Natural language processing

1 Introduction

The importance of objective scientific information is universally acknowledged. In practice, accessing, processing and comprehending relevant scientific documents is challenging, due to complex terminology and the potential lack of prior knowledge among readers. The CLEF 2024 SimpleText track aims at improving accessibility to scientific information for everyone both in terms of information

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14613, pp. 28–35, 2024.
https://doi.org/10.1007/978-3-031-56072-9_4

retrieval and natural language processing. The workshop at CLEF 2021 [3] and tracks at CLEF 2022-2023 [5,6] resulted in research community and test collections for improving access to scientific information for everyone. Specifically, test collections for retrieving relevant (and accessible) scientific text [11], for simplifying the language used in scientific documents without compromising the accuracy of the information [4], and for making complex concepts more understandable to a broader audience [2].

Scientific Text Simplification is different from traditional text simplification approaches focusing on lower literacy levels, for example making general text accessible to youth readers. Recent advances in IR and NLP hold the promise of removing some of the barriers to scientific information access.[1] The overall impact of CLEF SimpleText is to increase science literacy and broaden the audience of objective, scientific information.

The track's setup is based on the following pipeline: i) select the information to be included in a simplified summary; ii) improve the readability of the scientific text; iii) provide additional background knowledge for remaining difficult concepts; and iv) aggregate information from multiple articles. This results in the following four tasks [6]:

- ***Task 1: Content Selection*** *retrieving passages to include in a simplified summary.*
- ***Task 2: Complexity Spotting*** *identifying and explaining difficult concepts.*
- ***Task 3: Text Simplification*** *simplify scientific text.*
- ***Task 4: SOTA?*** *tracking the state-of-the-art in scholarly publications.*

In the rest of this paper, we will first reflect on the CLEF 2023 edition of the track in Sect. 2, and then provide a detailed description of each task of the CLEF 2024 edition in Sect. 3. We end with a discussion and conclusions in Sect. 4.

2 Results and Lessons from CLEF 2023 SimpleText

For the second year of running SimpleText as a track at CLEF 2023, 79 teams were registered [5]. Among them, 20 teams submitted 139 runs. For Task 1 (selecting passages/abstracts to include) [11], 39 runs were submitted by 5 teams. For Task 2 (identifying difficult terms) [2], we received 39 runs by 12 teams for subtask 2.1, and 29 runs by 10 teams for 2.2. For Task 3 (rewriting text) [4], a total of 32 submissions by 14 teams was made. The increase in active participation was encouraging.

For Task 1, we extended last year the scientific passage retrieval test collection, with a high pooling diversity, and reusable with limited pooling bias. Almost all submissions were based on neural rankers. Crossencoders and bi-encoders were popular approaches and turned out to be very effective. Promising results were observed for runs prioritizing credibility/complexity. This interesting feature can guide users to accessible content first, and more complex text

[1] A joined effort with Scholarly Document Processing https://sdproc.org/2024/.

later. In 2024, we will extend these qrels by increasing pooling depth and adding new subtopics and queries for the same set of popular science articles. We will also add supplementary labels on text complexity.

In the 2023 edition of Task 2, in addition to difficult term spotting (Task 2.1), we also asked participants to provide a definition or explanation of these concepts (Task 2.2). For the first task, both LLMs and traditional keyword extraction approaches performed well, but for the second task, LLMs outperformed traditional approaches. For Task 2.2, evaluation abbreviation or acronym expansion against ground truth is straightforward. For concepts, a set of reference sentences describing the concept was used, but many possible definitions or explanations may exist. In 2024, we will specifically evaluate the usefulness and difficulty of these explanations.

For Task 3, LLMs proved very effective in generating text simplifications, as well as for the highly complex scientific text used in the track's corpus. In general, larger models outperformed earlier models, and the limited training instances led to further improvement (but also to potential overfitting). As LLMs are used in generative mode, the analysis revealed varying degrees of hallucination where the models generated additional (and very plausible) extra content not warranted by the original input. Remarkably, this is ignored by standard evaluation measures (SARI, BLEU, ROUGE) based on text overlap with reference sentences. In 2024, we will introduce new evaluation measures that quantify unsupported content. In addition to the current sentence-based text simplification, we will also provide novel passage-based text simplification input and evaluation.

Our shared tasks are interconnected. The corpus used in Tasks 2 and 3 is based on abstracts in response to a popular science request in Task 1. In 2024, we will further expand the SimpleText test collections, provide additional evaluation measures. We will also introduce a new SOTA task aiming to generate a structured summary of scientific knowledge in multiple papers.

3 CLEF 2024 SimpleText Tasks

We will keep the three tasks from the 2023 edition and add a new one. We will reuse data constructed in previous editions with additional topics and additional automatic and manual labels. We will also emphasize automatic evaluation and training using the 2023 data.

3.1 Task 1: Retrieving Passages to Include in a Simplified Summary

Given a popular science article targeted to a general audience, this task aims at retrieving passages, which can help to understand this article, from a large corpus of academic abstracts and bibliographic metadata. Relevant passages should relate to any of the topics in the source article.

Data. We use popular science articles as a source for the types of topics the general public is interested in and as a validation of the reading level that is suitable for them. The main corpus is a large set of scientific abstracts plus

associated metadata covering the field of computer science and engineering. We reuse the collection of academic abstracts from the Citation Network Dataset (12th version released in 2020)[2] [12]. This collection was extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. It includes, in particular, 4,232,520 abstracts in English, published before 2020. Search requests are based on popular press articles targeted to a general audience, based on *The Guardian* and *Tech Xplore*. Each of these popular science articles represents a general topic that has to be analyzed to retrieve relevant scientific information from the corpus.

We provide the URLs to original articles, the title, and the textual content of each popular science article as a general topic. Each general topic was also enriched with one or more specific keyword queries manually extracted from their content, creating a familiar information retrieval task ranking passages or abstracts in response to a query. Available training data from 2023 includes 29 (train) and 34 (test) queries, with the later set having an extensive recall base due to the large number of submissions in 2023 [11]. In 2024, we will extend this test collection with additional test queries.

Evaluation. Topical relevance was evaluated last year with a 0–2 score on the relevance degree towards the content of the original article. In 2023, we provided an initial analysis of text complexity (based on readability measures) and authoritativeness (based on academic impact measures). In 2024, we plan to provide additional evaluation measures on both topical relevance and complexity/credibility. While these criteria can provide different levels of comparison between systems, we will continue to provide standard ranking scores based on NDCG.

3.2 Task 2: Identifying and Explaining Difficult Concepts

The goal of this task is to decide which concepts in scientific abstracts require explanation and contextualization in order to help a reader understand the scientific text. Since 2023, we ask participants to identify such concepts and to provide useful and understandable explanations for them. Thus, the task has two steps: i) to identify candidate terms in a given passage from a scientific abstract and set the level of difficulty of each term (easy or hard); ii) to provide a definition or an explanation or both only for the difficult (hard) terms.

Data. The corpus of Task 2 is based on the sentences in high-ranked abstracts to the requests of Task 1. New 2024 test data will be based on 116,763 sentences from the DBLP scientific abstracts used in Task 1.

Training data for the first step of the task, i.e. retrieving difficult terms, is based on the train and test data collected in 2023 [2]. The 2022 train data consists of 203 pairs sentence/term plus term definitions, and the 2022 test data consists of 5,142 distinct pairs sentence/term pooled from the participants' runs (1,262 distinct sentences).

[2] https://www.aminer.cn/citation.

Similarly, for the second step of the task, there is 2023 data available for training based on 1,000 ground truth definitions collected by Elsevier, and 5,000 mined abbreviations. The first set is extracted from a much larger corpus of full-text articles, extracted from books and articles published in ScienceDirect[3]. Moreover, there will be terminological definitions and explanations manually generated available for a subset of the training and test data used in Task 3 in 2023. A total of 175 documents and 893 sentences will be manually annotated. Finally, we encourage participants to train on existing datasets extracted from other resources such as the WCL dataset [10] to train the definition generation model, or use gazetteers, wikification resources as well as resources for abbreviation deciphering.

Evaluation. We will evaluate complex concept spotting in terms of their complexity and the detected concept spans [2]. We will automatically evaluate provided explanations by comparing them to references (e.g. ROUGE, cosine similarity, etc.). In addition, we will manually evaluate the provided explanations in terms of their usefulness with regard to a query as well as their complexity for a general audience. Note that the provided explanations can have different forms, e.g. abbreviation deciphering, examples, use cases, etc.

3.3 Task 3: Simplify Scientific Text

The goal of this task is to provide a simplified version of sentences extracted from scientific abstracts. Participants will be provided with the popular science articles and queries and matching abstracts of scientific papers, split into individual sentences.

Data. Task 3 uses the same corpus based on the sentences in high-ranked abstracts to the requests of Task 1. Our training data is a truly parallel corpus of directly simplified sentences coming from scientific abstracts from the DBLP Citation Network Dataset for *Computer Science* and Google Scholar and PubMed articles on *Health and Medicine*. Available training data from 2023 includes 648 sentences (train) and 245 sentences (test) from scientific abstracts plus manual simplifications [4]. These text passages were simplified either by master students in Technical Writing and Translation or by a domain expert (a computer scientist) and a professional translator (native English speaker) working together.

Other existing text simplification corpora used post-hoc aligned sentences [e.g., [13]. The SimpleText corpus contains 900 directly simplified sentences, and a useful addition to existing high-quality corpora like NEWSELA [13] (2,259 sentences). Our track is the first to focus on the simplification of scientific text with a much higher text complexity than news articles. In 2024, we will expand the training and evaluation data. In addition to sentence-level text simplification, we will provide passage-level input and reference simplifications, with the train and test data corresponding to 137 and 38 abstracts respectively.

[3] https://www.sciencedirect.com/.

Evaluation. In 2024, we will emphasize large-scale automatic evaluation measures (SARI, ROUGE, compression, readability) that provide a reusable test collection. This automatic evaluation will be supplemented with a detailed human evaluation of other aspects, essential for deeper analysis. As in 2023, we evaluate the complexity of the provided simplifications in terms of vocabulary and syntax as well as the errors (Incorrect syntax; Unresolved anaphora due to simplification; Unnecessary repetition/iteration; Spelling, typographic or punctuation errors) [4]. Almost all participants used generative models for text simplification, yet existing evaluation measures are blind to potential hallucinations with extra or distorted content [4]. In 2024, we will provide new evaluation measures that detect and quantify hallucinations in the output.

3.4 Task 4: Tracking the State-of-the-Art in Scholarly Publications

In Artificial Intelligence (AI), a common research objective is the development of new models that can report state-of-the-art (SOTA) performance. The reporting usually comprises four integral elements: Task, Dataset, Metric, and Score. These (Task, Dataset, Metric, Score) tuples coming from various AI research papers go on to power leaderboards in the community. Leaderboards, akin to scoreboards, traditionally curated by the community, are platforms displaying various AI model scores for specific tasks, datasets, and metrics. Examples of such platforms include the benchmarks feature on the Open Research Knowledge Graph and Papers with Code (PwC). Utilizing text mining techniques allows for a transition from the conventional community-based leaderboard curation to an automated text mining approach. Consequently, the goal of Task 4: SOTA? is to develop systems which given the full text of an AI paper, are capable of recognizing whether an incoming AI paper indeed reports model scores on benchmark datasets, and if so, to extract all pertinent (Task, Dataset, Metric, Score) tuples presented within the paper.

Data. The training and test datasets for this task are derived from community-curated (T, D, M, S) annotations for thousands of AI articles available on PwC (CC BY-SA). We will utilize the dataset obtained from our prior work, specifically the PwC source downloaded on May 10, 2021 [7,8], which comprised over 7,500 articles. These articles, originally sourced from arXiv under CC-BY licenses, are available in TEI XML format, each accompanied by one or more (T, D, M, S) annotations from PwC. While our previous work employed dataset splits for two-fold cross-validation experiments, for the SimpleText Task 4, we will establish new 70/30 train/test splits, providing approximately 5,000 annotated articles for participant training. A preliminary version of our training dataset can be accessed on Github https://github.com/jd-coderepos/sota.

The test set will strategically include only those articles with TDMs seen in the training set, creating a few-shot evaluation setting. Furthermore, in our subsequent research [9], we explored a zero-shot evaluation setting, wherein the dataset contained articles with at least one T, D, or M not seen in the model's training set. Thus in addition to the few-shot evaluation, we intend to introduce

a second evaluation setting for Task 4, evaluating models in a zero-shot context, for which a new test dataset will be created. Finally, ongoing efforts involve expanding the primary task corpus by incorporating approximately 1,500 articles into both the train and test sets that do not report leaderboards. These articles will be annotated with the *unknown* label. Consequently, systems developed in our shared task will have comprehensive applicability to any AI article, extracting (T, D, M, S) annotations for articles that contain them and assigning *unknown* for those that do not.

Evaluation. As discussed above, in Task 4 participant systems will be evaluated in the two evaluation settings. For **Few-shot** evaluation, trained systems will have to predict (T, D, M, S) annotations on a new collection of articles' full-text. The labels in the gold dataset will include only (T, D, M, S)'s seen at least once in training. For **Zero-shot** evaluation, the task is as above with a different collection of articles, which have (T, D, M, S) with unseen T, D, or M in the training set. In both settings, the standard recall, precision, and F-score metrics will be used to report scores to the participant systems.

4 Conclusions

This paper described the setup of the CLEF 2024 SimpleText track, which contains four interconnected tasks on scientific text summarization and simplification. Within the SimpleText track, we have already released extensive corpora and manually labeled data. First, a large corpus of over 4 million scientific abstracts that can be used for popular science. Second, scientific terms from sentences coming from scientific abstracts with manually attributed difficulty scores. Third, a parallel corpus of manually simplified sentences from scientific literature. Fourth, a parallel corpus of sentences with different types of information distortion and simplification level. Please visit the SimpleText website (http://simpletext-project.com) for more details on the track.

Acknowledgments. This track would not have been possible without the great support of numerous individuals. We want to thank in particular the colleagues and the students who participated in data construction, evaluation and reviewing. We also thank the MaDICS (https://www.madics.fr/ateliers/simpletext/) research group and the French National Research Agency (project *ANR-22-CE23-0019-01*). SimpleText's SOTA Task is jointly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number: NFDI4DataScience (460234259) and the German BMBF project SCINEXT (01lS22070).

References

1. Aliannejadi, M., Faggioli, G., Ferro, N., Vlachos, M. (eds.): Working Notes of CLEF 2023: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, vol. 3497, CEUR-WS.org (2023). http://ceur-ws.org/Vol-3497
2. Ermakova, L., Azarbonyad, H., Bertin, S., Augereau, O.: Overview of the CLEF 2023 SimpleText Task 2: difficult concept identification and explanation. In: [1]. https://ceur-ws.org/Vol-3497/paper-239.pdf
3. Ermakova, L., et al.: Text Simplification for Scientific Information Access: CLEF 2021 SimpleText Workshop. In: Advances in Information Retrieval - 43nd European Conference on IR Research, ECIR 2021, Lucca, Italy, March 28 - April 1, 2021, Proc., Lucca, Italy (2021)
4. Ermakova, L., Bertin, S., McCombie, H., Kamps, J.: Overview of the CLEF 2023 SimpleText Task 3: Scientific text simplification. In: [1]. https://ceur-ws.org/Vol-3497/paper-240.pdf
5. Ermakova, L., SanJuan, E., Huet, S., Azarbonyad, H., Augereau, O., Kamps, J.: Overview of the CLEF 2023 SimpleText Lab: automatic simplification of scientific texts. In: Arampatzis, A., et al. (eds.) CLEF'23: Proceedings of the Fourteenth International Conference of the CLEF Association. LNCS. Springer (2023). https://doi.org/10.1007/978-3-031-42448-9_30
6. Ermakova, L., et al.: Overview of the CLEF 2022 SimpleText lab: automatic simplification of scientific texts. In: Barrón-Cedeño, A., et al. (eds.) CLEF'22: Proceedings of the Thirteenth International Conference of the CLEF Association. LNCS. Springer (2022)
7. Kabongo, S., D'Souza, J., Auer, S.: Automated mining of leaderboards for empirical ai research. In: Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23, pp. 453–470. Springer (2021)
8. Kabongo, S., D'Souza, J., Auer, S.: Orkg-leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. arXiv preprint arXiv:2305.11068 (2023)
9. Kabongo, S., D'Souza, J., Auer, S.: Zero-shot entailment of leaderboards for empirical ai research. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2023 (2023)
10. Navigli, R., Velardi, P.: Learning word-class lattices for definition and hypernym extraction. In: ACL, pp. 1318–1327 (2010)
11. SanJuan, E., Huet, S., Kamps, J., Ermakova, L.: Overview of the CLEF 2023 simpletext task 1: passage selection for a simplified summary. In: [1]. https://ceur-ws.org/Vol-3497/paper-238.pdf
12. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: KDD'08, pp. 990–998 (2008)
13. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: new data can help. Trans. ACL **3**, 283–297 (2015). ISSN 2307–387X. https://www.mitpressjournals.org/doi/abs/10.1162/tacl_a_00139